

**MASCHINENLESBARE WÖRTERBÜCHER ALS INSTRUMENT
FÜR SYNTAKTISCHE UND SEMANTISCHE ANALYSEN
(EIN VERGLEICH)**

Abhandlung

zur Erlangung der Doktorwürde
der Philosophischen Fakultät I
der
Universität Zürich

vorgelegt von
Hans Huonker
von Zürich

Angenommen auf Antrag von Herrn Prof. Dr. Harald Burger

Zürich 1987

ADAG Administration & Druck AG

MASCHINENLESBARE WÖRTERBÜCHER ALS INSTRUMENT FÜR SYNTAKTISCHE UND SEMANTISCHE ANALYSEN

(EIN VERGLEICH)

Abhandlung

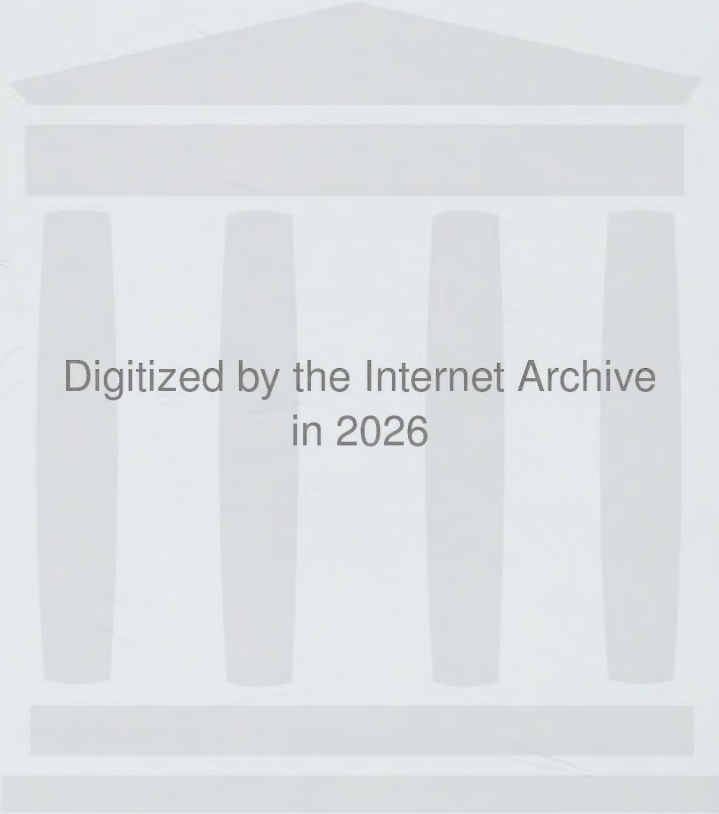
zur Erlangung der Doktorwürde
der Philosophischen Fakultät I
der
Universität Zürich

vorgelegt von
Hans Huonker
von Zürich

Angenommen auf Antrag von Herrn Prof. Dr. Harald Burger

Zürich 1987

ADAG Administration & Druck AG



Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41507201_0045

Bei den in dieser Arbeit untersuchten Wörterbüchern handelt es sich um maschinenlesbare Lexika. Probleme maschinenlesbarer Wörterbücher aus dem Bereich der maschinellen Übersetzung werden nicht behandelt. Die hier untersuchten Lexika sind mit einer Ausnahme einsprachige Wörterbücher und dienen der Explikation morphologischer, syntaktischer, semantischer und anderer Informationen, welche in Sprachdaten verarbeitenden Systemen verwendet werden.

Schwerpunkt der Untersuchung ist der linguistische Aspekt derartiger Lexikonsysteme, nicht der informatische. Dazu wird in Teil 1 der Arbeit der lexikographische Zusammenhang entwickelt. Die informatische Seite wird hauptsächlich vom Anwender-Standpunkt aus betrachtet. Aus linguistischer Perspektive werden, abgesehen von allgemeinen lexikographischen Reflexionen in Bezug auf maschinenlesbare Wörterbücher, die Probleme der lexikalischen Semantik erörtert, die sich bei Formalisierungen und Operationalisierungen ergeben.

Im 2. und im 3. Teil der Arbeit werden Resultate einer empirischen Untersuchung präsentiert, kommentiert und ausgewertet. Die Daten basieren auf einer Umfrage aus den Jahren 1984/85 und beziehen sich auf die Lexikonsysteme von 6 Projekten, drei davon aus dem Umfeld 'Artificial Intelligence', zwei aus dem Gebiet der 'Linguistischen Datenverarbeitung' und eines aus der industriellen Praxis (Terminologie).

Weil sich die elektronische Datenverarbeitung sehr schnell entwickelt, besteht bei einer empirischen Untersuchung die Gefahr, dass die Daten rasch veralten. Um diesen Trend zu dokumentieren, wird an einem der 6 Lexikonsysteme exemplarisch der diachronische Entwicklungsaspekt ausgeleuchtet.

*

An dieser Stelle bedanke ich mich bei all jenen, die mir bei dieser Arbeit halfen, zuerst bei Herrn Professor Dr. Harald Burger, der bereit war, eine Arbeit zum Thema "Maschinenlesbare Lexika" als Inhalt meiner Dissertation zu akzeptieren und zu unterstützen. Danken möchte ich auch Herrn Doktor Moscheh Mresse, Privatdozent für Informatik, der sich die Zeit nahm, mein Manuskript zu lesen, zu kommentieren, kritische Stellen ausführlich mit mir zu besprechen und ein Gutachten abzugeben. Grossen Dank bin ich sodann Raimund Drewek schuldig, der trotz starker Arbeitsbelastung immer wieder Zeit fand, sich mit mir freundschaftlich und ausserordentlich kompetent über theoretische und praktische Aspekte der computergestützten Lexikographie zu unterhalten und mir in entscheidenden Punkten sehr wertvolle Hinweise und Anregungen zu geben.

Danken möchte ich auch allen jenen, die sich die Mühe genommen haben, den von mir verschickten Fragebogen auszufüllen und bereit waren, mir bei Rückfragen in offener und kollegialer Atmosphäre

weitere detaillierte Auskünfte zu erteilen.

Im weiteren gebührt mein Dank den Mitgliedern des Programmkomitees der Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung, welche mir im Februar 1986 die Gelegenheit gaben, im Rahmen der Tagung zum Thema "Reflexionen zum Begriff 'Wortdatenbank' in der Linguistischen Datenverarbeitung" zu sprechen.

Dank verdienen auch meine Freunde Georg Brun und Thomas Schnurrenberger von der Universität Zürich für eingehende fachliche Gespräche und sorgfältiges, kritisches Durchlesen des Manuskripts.

Zum Schluss komme ich zum Dank an meine Eltern, meine Frau und meinen kleinen Sohn, denen ich Hilfe, Unterstützung und nicht zuletzt Auflockerung bei einer langandauernden Arbeit verdanke.

ABSTRACT

Based on empirical data and lexicological theory I investigate various problems in the field of 'formal semantics in machine readable dictionaries'. The empirical data is about the dictionary systems of 6 natural language processing applications. Five of them are monolingual systems (german) and one of them is a large multilingual dictionary. The focus of the investigation is on the comparison between the relevant semantic categories in those dictionaries.

TABLE OF CONTENTS

Vorwort	3
Abstract	5
List of Illustrations	7
TEIL 1 Lexikographische Probleme.	10
1.1 * Einsprachige, deutsche maschinenlesbare Wörterbücher.	10
1.1.1 Formalisierbarkeit.	11
1.1.2 Implementation.	23
1.1.3 Adäquatheit der Repräsentation.	38
1.2 * Lemma und Eintrag.	48
1.2.1 Eindeutigkeit.	49
1.2.2 Disambigulierung.	51
1.2.3 Informationslinguistisch fundierter Eintrag.	52
1.3 * Verflechtungen zwischen der Mikro- und der Makrostruktur.	55
1.3.1 Analysierbare Teile des Lexikons.	57
1.3.2 Knowledge Engineering.	59
1.3.3 Lexikographische Adäquatheit.	60
TEIL 2 Empirische Erkenntnisse aus einer Umfrage.	61
2.1 * Methode und Ziel der Umfrage.	61
2.1.1 Auswahl der Projekte.	64
2.1.2 Die Formulierung der Fragen.	67
2.1.3 Echo, Präsentation, Auswertung.	68
2.2 * Die Antworten nach Projekten geordnet.	68
2.2.1 Das Lexikon des Systems HAM-ANS.	68
2.2.2 Das Lexikon des Systems KIT.	79
2.2.3 Das Lexikon des Systems EVAR.	82
2.2.4 Das Lexikon SADAW.	93

2.2.5	Das Lexikonsystem ALEXSYS.	100
2.2.6	Das Lexikon TEAM.	111
TEIL 3	Vergleichende Auswertung.	121
3.1 *	Inhaltliche Gruppierung der Antworten.	121
3.1.1	Linguistische Konzeption.	123
3.1.2	Datengrundlage.	148
3.1.3	Implementation.	150
3.1.4	Quantitative Merkmale und Restriktionen.	154
3.1.5	Umfeld.	155
3.1.6	Hardware.	156
3.2 *	Vergleich semantischer Kategorien.	157
3.2.1	Identifizierung der semantischen Kategorien	158
3.2.2	Einbeziehung des semantischen Bezugssystems.	167
3.2.3	Vergleich ableitbarer semantischen Konzeptionen.	170
3.3 *	Weitere Auswertungen.	179
3.3.1	Synchronische Aspekte der Umfrage.	179
3.3.2	Diachronische Aspekte.	182
3.3.3	Abschätzung des Erkenntniswerts der Untersuchung.	196
Anmerkungen		203
Anmerkungen zu Teil 1.		203
Anmerkungen zu Teil 2.		208
Anmerkungen zu Teil 3.		213
Verwendete Literatur		218
Lebenslauf		232
Anhang		233

LIST OF ILLUSTRATIONS

Figure 1.	'Frühes semantisches Netzwerk'	21
Figure 2.	'Morphosyntaktische Angaben (ALEXSYS)'	28
Figure 3.	'Ausschnitt Lexikoneintrag (EVAR)'	39
Figure 4.	'Analyseprozesse und Wissensquellen'	46
Figure 5.	'Umfang ausgewählter Lexika'	65
Figure 6.	'SUTRA-Wortklassen (HAM-ANS)'	75
Figure 7.	'Forts. Wortklassen (HAM-ANS)'	76
Figure 8.	'Relationen im Wortlexikon (HAM-ANS)'	77
Figure 9.	'Syntax-Modul (EVAR)'	85
Figure 10.	'Wortklassen (EVAR)'	90
Figure 11.	'Semantik-Eintrag (EVAR)'	92
Figure 12.	'Analysewortklassen (SADAW)'	98
Figure 13.	'Forts. Analysewortklassen (SADAW)'	99
Figure 14.	'Datenbasis des kumulierten Lexikons'	101
Figure 15.	'Informationsklassen (ALEXSYS)'	108
Figure 16.	'Wortklassen (ALEXSYS)'	110
Figure 17.	'Datenbankgeneriertes technisches Wörterbuch'	113
Figure 18.	'Kategorienschema (TEAM)'	119
Figure 19.	'Kategorienschema (TEAM)'	120
Figure 20.	'Wortartenverband (KIT)'	128
Figure 21.	'Wortklassen (EVAR)'	130
Figure 22.	'Forts. Wortklassen (EVAR)'	132
Figure 23.	'Wortexperte (KIT)'	137
Figure 24.	'Wortstammexperte (KIT)'	142
Figure 25.	'Verbstammeintrag (SADAW)'	144
Figure 26.	'Kasusrahmen (HAM-ANS)'	160
Figure 27.	'Sortenverband (KIT)'	163
Figure 28.	'Onomasiologisches Strukturbild'	177
Figure 29.	'Kasusrahmen (EVAR)'	189
Figure 30.	'Semantik-Modul (EVAR)'	191

TEIL 1 LEXIKOGRAPHISCHE PROBLEME.

Der Hauptakzent der vorliegenden Arbeit liegt auf der Analyse der Eintragsstruktur deutscher maschinenlesbarer Wörterbücher. U.a. werden die lexiko-semantischen Teile der Eintragsstruktur untersucht.

Unter Eintragsstruktur werden dabei im Sinne einer Arbeitshypothese die Abfolgeschemata von explizit kategorisierten Eintrags-elementen eines maschinenlesbaren Wörterbuchs verstanden. Solche Abfolgeschemata findet man auch in gedruckten Wörterbüchern. Dazu ein Beispiel aus dem Deutschen Wörterbuch von G. Wahrig: Das Abfolgeschema besteht hier aus vier Kategorien:

gleit-si-cher <Adj.> nicht gleitend

Suchwort, Trennstellen, Wortklasse, Paraphrase.

In den folgenden Kapiteln geht es aber um die Abfolgeschemata in maschinenlesbaren Wörterbüchern.

1.1 * EINSPRACHIGE, DEUTSCHE MASCHINENLESBARE WÖRTERBÜCHER.

Bei der Verwendung des Begriffs "maschinenlesbares Wörterbuch" ist zu unterscheiden, ob es sich lediglich um die Implementation eines konventionellen, vormals gedruckten Wörterbuchs handelt, oder aber um die strukturelle Neukonzeption eines Lexikonsystems im Rahmen eines Sprachdaten verarbeitenden Systems. Allerdings gibt es auch den zwischen diesen beiden Ansätzen liegenden Typ von maschinenlesbarem Wörterbuch:

Intention der Speicherung eines oder mehrerer Wörterbücher auf ein File-System oder eine Datenbank ist dann eine flexiblere Handhabung von Lexika im Editionsprozess. Maschinell erfasste Wörterbücher können zur rationellen Generierung neuer oder revidierter Ausgaben genutzt werden. Mittels Invertierung, Selektion und Kombination kann Umfang und Inhalt einer Ausgabe in relativ kurzer Zeit geändert werden. In (Drewek85, p.34) wird ein derartiges Konzept für einen Verlag vorgestellt, das auf einem DBMS und einem Netzwerkprozessor basiert.

Abgesehen vom Verlagswesen evoziert der Begriff "maschinenlesbares Wörterbuch" u.a. die folgenden Zusammenhänge:

- Wörterbücher für Parsersysteme.

Die Lexikoneinträge sollen, abgesehen von den Wortklassenangaben, die Gesamtheit der Irregularitäten des zu parsenden Sprachsubsets

enthalten, während Regelhaftes durch die Grammatik im Parserprogramm repräsentiert wird.

" Damit ist hier also eine Zweiteilung des Sprachsystems in Regelsystem (Regularitäten enthaltend) und Lexikon (Irregularitäten enthaltend) vorgenommen ".
(Zimmermann72, p. 7)

Zimmermann bezieht sich an dieser Stelle auf die Theorie der TG von Chomsky (aspects). An anderer Stelle mit Bezug auf Parser und Lexikon schreibt Zimmermann:

"Die Lexikoneintragung ist in Aufbau und Inhalt abhängig von dem Gebrauch, der im Analyseprozess von ihr gemacht werden soll. Je anspruchsvoller dabei der Parser ist, um so differenzierter ist ihr Informationsgehalt."
(Zimmermann72, p. 25)

- Maschinenlesbare Wörterbücher werden auch in der Maschinellen Übersetzung eingesetzt, sowohl bei maschinenunterstützten Verfahren als auch bei automatisierten. Während jeder der drei Phasen Analyse, Transfer und Synthese werden maschinenlesbare Wörterbücher benötigt.

- Als ein der Maschinellen Übersetzung benachbartes Einsatzgebiet von maschinenlesbaren Wörterbüchern seien Terminologiedatenbanken erwähnt. Diese können ein- oder mehrsprachig sein und dienen neben der terminologischen Normung auch der Unterstützung von technischen Übersetzern und Übersetzerinnen.

-Auf die in TEIL 2 dieser Arbeit untersuchten Sprachdaten verarbeitenden Systeme und ihre Lexikonkomponenten sei hier ebenfalls verwiesen.

1.1.1 Formalisierbarkeit.

Ausgangspunkt der Überlegungen zur Formalisierbarkeit der Eintragsstruktur maschinenlesbarer Wörterbücher ist der bereits eingeführte Begriff des Abfolgeschemas.

"Sind alle Elemente eines Eintrags (s. DIN2333) [**1] explizit kategorisiert, so nennt man alle in einem Wörterbuch verwendeten Abfolgeschemata von Kategorien die Struktur des Wörterbucheintrags.

Ein Wörterbucheintrag wie zum Beispiel

NOTE, sub.f., 1.Musikzeichen, 2.Zensur, 3.Geldschein,
4.Regierungsmittelung

hat eine Struktur, die als Abfolge der Kategorien

(Stichwort + Wortklasse + Genus + (Bedeutung)*)

beschreibbar ist. Der an die Kategorie (Bedeutung) angefügte Stern symbolisiert, dass eine beliebige Wiederholung möglich ist."
(Drewek82, p.9)

Durch systematische Variation des Abfolgeschemas lassen sich Eintragsstrukturen von unterschiedlicher Komplexität erzeugen.

Abstrakte Formalisierungen.

Die so erzeugten abstrakten Formalisierungen der Eintragsstruktur maschinenlesbarer Wörterbücher, angeordnet in aufsteigenden Komplexitätsstufen, sind:

- a) Wortlisten.
- b) Zweiteiliges Abfolgeschema.
- c) Variable Abfolgeschemata.
- d) Tabellen.
- e) Baumstrukturen.
- f) Syntaktisch organisierte Strukturen.
- g) Wissensrepräsentationsformalismen.

Im folgenden werden diese Stufen systematisch entwickelt.

a) Wortlisten.

Die einfachste abstrakte Formalisierung der Eintragsstruktur ist die Wortliste. Der Begriff Wort soll auf dieser Stufe der formalen Abstraktion nicht weiter differenziert werden. Wortlisten bestehen aus der Summe von ein-elementigen Abfolgeschemata. Die hier implizierte Voraussetzung, dass ein Abfolgeschema aus lediglich einem Element bestehen kann, dehnt den Begriff "Abfolgeschema" zwar über seinen üblichen Gebrauch aus, doch eignet sich diese Abstraktion, um auch einfachere, praktische lexikographische Tätigkeiten systematisch im Kontext "Wörterbuch" zu beschreiben. Über Wortlisten können bereits viele elementare lexikographische Tätigkeiten durchgeführt werden: man kann die Listen sortieren, vergleichen, verkürzen, erweitern uam.

Diese Tätigkeiten sind nicht zu unterschätzen und bilden die Basis der Wörterbucharbeit. Mit Hilfe des Computers können diese Operationen über sehr umfangreichen Listen innerhalb kurzer Zeit durchgeführt werden. Beispiele für derartige elementare Wortlisten sind Wortformenlisten, Vokabulare und Ranglisten. Wortformenlisten können nach der über einem Alphabet definierten Sortierfolge geordnet werden, Ranglisten nach der Häufigkeit des Auftretens der Wortform.

Sind die Einträge in solchen Wortlisten mit mehr Information als nur der Sortierfolge und der Orthographie versehen, verändert sich ihre formale Struktur. Wenn einem Worteintrag in einem Wortformenindex die Referenz auf die Fundstelle zugeordnet ist, oder wenn in einer Frequenzliste zu jedem Type die Häufigkeit vermerkt ist, dann handelt es sich nicht mehr um elementare Wortlisten.

b) Zweiteiliges Abfolgeschema.

Fügt man zu jedem Worteintrag einen Informationsteil hinzu, kann man bereits von einem Wörterbuch (oder Glossar) sprechen. Das in Wortlisten realisierte ein-elementige Abfolgeschema erweitert sich zu einem zweiteiligen Schema, bestehend aus Worteintrag und Informationsteil. Im einfachsten Fall besteht der Informationsteil nur aus einer einzigen explizit erwähnten Kategorie, z.B. aus dem dem Wort (oder der Wortform) zugeordneten Wortart. Derartig reduzierte Wörterbücher werden häufig zu Testzwecken für den Lexikonteil eines syntaktischen Parsers eingesetzt. Der Parser erkennt ein Satzsegment als Wort, vergleicht es mit den Lexikoneinträgen, liest die Wortklassenangabe des passenden Wortes und setzt mit Hilfe dieser externen Information die auf dem internen grammatikalischen Regelapparat basierende Satzanalyse fort.

Der Informationsteil eines einfachen zweiteiligen Wörterbuchs muss aber nicht eindeutig kategorisiert sein. Analog zum Wörterbuchartikel eines gedruckten Wörterbuchs kann dieser Teil des Abfolgeschemas eine beliebige Menge Text sein. Obwohl in gedruckten Wörterbüchern der Textaufbau eines Wörterbuchartikels bei weitem nicht beliebig ist, so ist er doch auch nicht immer eindeutig. Es gibt in einem Wörterbuchartikel zwar eine Gruppierung in verschiedene Informationskategorien, aber es sind nie alle Kategorien in gleicher Form bei allen Einträgen realisiert.

Anstatt diese Ordnung für ein maschinenlesbares Wörterbuch zu rekonstruieren, kann man den gesamten Informationsteil abstrakt als eine Kategorie verstehen, nämlich als Textmenge. Die Suche im Wörterbuch gestaltet sich über das sortierte Stichwort (Lemma) immer noch sehr einfach und effizient. Wie dann der gefundene Text abgearbeitet wird, ist eine andere Frage. Die Suche kann aber auch direkt in der Textmenge des Informationsteils durchgeführt werden. Man verfährt dann nach einer der Techniken, die im Information Retrieval über Dokumenten bei der Freitextsuche angewendet werden. (Vergl. Anm. [**5], Teil 1).

c) Variable Abfolgeschemata.

Durch explizite Untergliederung des Informationsteils in mehrere Informationskategorien erhält man eine abstrakte Struktur, die derjenigen eines gedruckten Wörterbuchs auf den ersten Blick recht ähnlich zu sein scheint. Eine Bedingung ist dabei, dass es verschiedene Arten von Abfolgeschemata geben darf. Diese Unterschiede können sich auf die Kategorienart und die Kategorienzahl beziehen. So kann z.B. jeder Wortart ein eigenes Abfolgeschema zugeordnet werden.

Bei dieser Art der Formalisierung der Eintragsstruktur sind viele Varianten möglich: Ein Abfolgeschema muss nicht vollständig realisiert werden. So muss z.B. nicht bei allen Substantiven die Kategorie "Umlaut" realisiert werden, wenn es keinen hat. Einzelne Kategorien können sich auch wiederholen: So besteht etwa die Kategorie "Paraphrase" bei einigen Wörtern nur aus einem Wort, bei anderen aber aus mehreren Ausdrücken. Damit ist auch gesagt, dass gleichartige Kategorien verschiedene Längen aufweisen dürfen, was sich dann ebenso auf die Länge gleichartiger Ablaufschemata auswirkt.

Mit Record-Strukturen können solche Informationstypen auf dem Computer nachgebildet und in Files gespeichert werden. Anschaulich gesprochen entspricht ein solches Wörterbuch etwa einer gut organisierten Kartei. Bei der EDV-mässigen Organisation einer solchen "Kartei" stellt sich ein grosser Vorteil ein: Wörterbuchinformationen können nicht nur über den Suchbegriff "Stichwort" gesucht werden, sondern über jede beliebige Kategorie. Für jede Kategorie kann eine Sortierfolge definiert werden. Rückläufige Wörterbücher sind z.B. nach der morphologischen Kategorie "Endung" sortiert.

Unter (1.1.2(c)) wird auf die Problematik der Implementation variabler Abfolgeschemata mit Recordstrukturen eingegangen und auf Alternativen verwiesen.

d) Tabellen.

Durch geeignete Restriktionen können aus variablen Abfolgeschemata Tabellen erzeugt werden. Der einfachere Fall tritt auf, wenn in jedem Abfolgeschema eine konstante Zahl von Elementen vorkommt und nur ein Schema zugelassen ist. Ein derart strukturiertes Wörterbuch gleicht einer Matrix, bestehend aus Zeilen von identischen Abfolgeschemata und Spalten von gleichen Kategorien. (D.h. die Tabelle weist ein einziges, starres Abfolgeschema auf) Theoretisch eignet sich diese Form auch für komplexe Wortschatzdarstellungen: Es können zu sämtlichen Wortarten sämtliche Kategorien gebildet und in einem grossen Abfolgeschema erfasst werden. Einträge für die wortartenspezifischen Informationen werden dann nur an denjenigen Stellen der Tabelle eingefügt, wo im Einheitsabfolgeschema eine relevante Kategorie auftritt. Die Kategorien werden in einem Codebook definiert (Typ, Wertebereich, Grösse).

Die tabellarische Darstellung des codierten Wortschatzes mit Matrizen lässt sich auch durch systematisches Aufspalten der Tabelle in kleinere Tabellen erreichen. Jede Tabelle hat dann ihr eigenes Abfolgeschema. So können die Tabellen z.B. nach Wortarten aufgebaut werden: Für jede Wortart eine Tabelle mit den dieser Wortart eigenen Kategorien im Abfolgeschema. Geht man noch einen Schritt weiter, so kann man auch innerhalb der Wortarten Informationen nach ihrer logischen Zusammengehörigkeit gruppieren, z.B. phonetische, morphologische, syntaktische, semantische, etymologische und pragmatische Kategoriengruppen.

Allerdings handelt man sich mit dem Gewinn an Übersichtlichkeit ein Problem ein, nämlich die verschiedenen Tabellen beim Lesen, Hinzufügen oder Verändern von Information zu einem Wort richtig verbinden zu müssen. Dieses Problem kann aber im Rahmen relationaler Datenbanken mittels geeigneter Schlüsselzuteilung auf die normalisierten Tabellen gelöst werden.

e) Baumstruktur.

Angelpunkt der Überlegungen zur Formalisierbarkeit von Wörterbucheinträgen und der Eintragsstruktur bildete bei den oben erwähnten Varianten (a-d) der Begriff eines linearen Abfolgeschemas. Das Abfolgeschema eines Eintrags braucht aber nicht linearer Art zu sein. Hierarchische Eintragsmuster entsprechen dem traditionellen, praxisorientierten lexikographischen Verständnis

eines Wörterbuchartikels eher als die tabellarische Verflachung dieses Musters.

"Die Ordnung - besser Anordnung - zeigt sich an der geschriebenen Oberfläche als lineare, doch wissen Lexikographen, dass jedem Eintrag Hierarchien zugrunde liegen. Die phonetische Umschrift gehört enger zum Stichwort, ebenso die Angaben zur Flexion. Eine Stilangabe warnt vor Gebrauch einer bestimmten Phrase in vornehmen Kreisen. Oder: zu einem Stichwort lassen sich mehrere Bedeutungen übersetzen. Deshalb ist es berechtigt anzunehmen, dass jedem Eintrag eine Baumstruktur zugewiesen werden kann, die solche Hierarchien beschreibt. Dem Linguisten kommt dies als Parsing-Problem wohlbekannt vor."
(Drewek85, p.36)

Diese Erkenntnis kann nun auf mindestens zwei Arten genutzt werden:

1. Bei der Analyse, also dem Parsen, bereits existierender, traditioneller Wörterbücher und
2. Beim Design neuer, maschinenlesbarer Wörterbücher durch Normierung der Eintragsgenerierung mittels einer Wörterbuchsyntax [**2].

Es wäre unvollständig, die Wörterbuchsyntax nur als Problematik der Eintragungssyntax zu verstehen, denn auch die Einträge sind nicht nur linear, bzw. sequentiell gereiht, sondern vielfältig verknüpft. (Es ist angebracht, zwischen der Syntax der Mikrostruktur und der Makrostruktur zu unterscheiden; vergl. Kapitel (1.3.*))

"Ein typisches Phänomen auf der Textebene bilden die Verweise, die eine Relation zwischen Einträgen oder Teilen davon aufbauen. Ein Thesaurus ist daher als weitgehend textuell verflochtenes Gebilde einzustufen. Nicht zu vergessen sind Benutzeranleitung, Werbetexte und Platzfüller, die als spezielle Subtexte in ein Wörterbuch integriert sein können.

Die gezielte linguistische Durchleuchtung der Lexikographie zeigt eigentlich nicht unerwartete Parallelen zur Entstehung von Texten natürlicher Sprache. Das Spezifische der Textsorte "Wörterbuch" ist in der Verwendung spezialisierter lexikalischer Kategorien und in einer deutlich ärmeren Syntax zu sehen. Dass Wörterbücher Texte über die natürliche Sprache selbst sind - die sogar mit der Intention abgefasst wurden, die verzeichnete Sprache (nicht nur jedoch) aus dem Wörterbuch zu lernen, sollte eigentlich keine Verwirrung stiften."
(Drewek85, p.37)

Mit diesen Beobachtungen an der Makrostruktur von Wörterbüchern ist nicht impliziert, dass bei der Verknüpfung der Einträge nur hierarchische Baumstrukturen gebildet werden. Generalisierend formuliert, ist es angebracht, von der Textgrammatik eines Wörterbuchs zu sprechen.

f) Syntaktisch organisierte Strukturen.

Tabellen und Recordstrukturen implizieren, dass Informationen im Wörterbuch hauptsächlich aus Elementen bestehen, die sich mit relativ wenigen Abhängigkeitsregeln effizient gruppieren lassen. Zu einem Teil ist das auch richtig, weil mit wenig Regeln schon

relativ viele Elemente gruppiert werden können. So lässt sich mit der Zuordnung von wortartabhängigen Kategorien zur entsprechenden Wortart schon eine dem Wörterbuch angenäherte Struktur nachbilden.

Geht man aber davon aus, dass ein Wörterbuch syntaktisch komplexer strukturiert ist, (was die oben zitierten Beobachtungen vermuten lassen), dann kann ein ganzer Regelapparat zu einer Wörterbuchsyntaktik aufgebaut werden. Ein Ausschnitt für die oberste Stufe eines derartigen Apparats sieht für das Lexikon des Systems EVAR (vergl. Kapitel 2.2.3) folgendermassen aus:

```
wort ::= [ WORT string
          WORTNR integer
          AUSSPRACHE ( phonetisches_Wort + )

          SYNTAX-TEIL
            ( wortartbeschreibung + ) ]

wortartbeschreibung ::=
  [ WORTARTNR integer.integer
    WORTART wortart
    morpho_syntaktische_beschreibung ]

morpho_syntaktische_beschreibung ::=
  NORMALFORM ( normalform ) + )
  { wortart_abhängige_syntaxbeschreibung,
    flektierbare_wortart, flekt.Form }
  ;
  { wortart_abhängige_syntaxbeschreibung,
    wortart_mit_semantik, NF }
  SEMANTIK-TEIL
    ( bedeutungsbeschreibung + )
  ;
  { wortart_abhängige_syntaxbeschreibung,
    wortart_ohne_semantik, NF }
```

```
normalform ::=
  [ WORT string
    WORTARTNR ( integer.integer + )
    { VERBPRAEFIX
      [ WORT string
        WORTARTNR integer.integer ] } 0.1
```

```
bedeutungsbeschreibung ::=
  [ BEDEUTUNGSNR integer.integer.integer
    BEDEUTUNG " string "
    { wortart_abhängige_semantikbeschreibung,
      wortart } ]
```

(Ehrlich86,p.26)

Diese Syntaxnotation ist so zu lesen:

```
[ ... ] ==>Beschreibung durch Attribut-Wert-Paare
( ... ) ==>Liste von Alternativen
```

Attribut-Wert-Paar:

"Behandlung der Stichwörter" wird u.a. erklärt, wie ein Artikel aufgebaut ist, wenn ein Stichwort mehrere Bedeutungen hat. So machen etwa römische Ziffern darauf aufmerksam, dass entweder

- a) verschiedene Wortarten vorliegen,
- b) die "einzelnen Bedeutungen inhaltlich ganz verschieden" sind
- c) es sich um Homographen handelt oder
- d) Homonymie vorliegt.

In einer formalisierten Syntax, wie sie in einem maschinenlesbaren Wörterbuch vorliegt, ist diese Syntaxexplikation nicht genügend genau, für den menschlichen Duden-Benutzer in den meisten Fällen hingegen schon.

Syntax-Definitionen in maschinenlesbaren Wörterbüchern dienen nicht nur der Strukturierung der Einträge und ihrer Verknüpfung, sondern bilden auch die Voraussetzung für eine allfällige Compilation. Genauso wie die Syntax-Definition einer Programmiersprache die Voraussetzung für die Compilation der in dieser Sprache verfassten Programme ist, ist das nach einer strikten Syntax verfasste Wörterbuch compilationsfähig. Dabei lässt sich Wörterbuchsyntaktik zwar weniger in der Nähe von Syntax-Definitionen prozeduraler Programmiersprachen ansiedeln, dafür umso eher mit der Syntax einer "Datenbank" vergleichen. Dabei muss differenziert werden, worauf sich der Begriff "Syntax" bezieht:

- Syntax der DML (Data Manipulation Language),
- Syntax der DDL (Data Definition Language) oder
- Syntax der für die gegebene Datenbank geltenden Konsistenzprüfungen. [**3]

Explizit und systematisch wird dieser Ansatz zur Definition eines Wörterbuchs mit Hilfe einer Datenbank in (Domenig86) anlässlich einer Sichtung und Modifizierung des 2-Ebenen-Modells von (Koskenniemi83) im Hinblick auf Verwendbarkeit im Eurotra-Projekt entwickelt. (Domenig86, p.129-185). (vergl. auch Kap. 1.1.2(f)).

g) Wissensrepräsentationsformalismen.

Formal können Wörterbücher nicht nur als syntaktisch organisierte Gebilde, sondern auch als Wissensrepräsentation gesehen werden. Diese Perspektive drängt sich bei Bedeutungswörterbüchern geradezu auf. Oft ist es so, dass die Grenze zwischen enzyklopädischem Lexikon und Bedeutungswörterbuch fließend ist. In Bedeutungswörterbüchern sind z. T. recht exotische, aber unter Umständen wissenswerte Wörter manchmal in enzyklopädischer Breite erklärt, und im Grossen Brockhaus ist bei vielen Stichwörtern etwas über ihre sprachliche Bedeutung zu erfahren.

Syntaktisch strukturierte Abfolgeschemata (Wörterbuchartikel) und deren textgrammatische Verknüpfung (Makrostruktur des Wörterbuchs) lassen sich als komplexe Wissensbasen interpretieren. Diese Idee ist nicht neu, denn bereits M.R. Quillian orientierte seine ersten semantischen Netzwerkmodelle nicht so sehr an der Funktionsweise des menschlichen Gehirns, als an Wörterbüchern. In (Wettler80, p.39ff) wird zwar darauf hingewiesen, dass diese ersten Modelle aus den Jahren 1967, 1968, 1969 [**4] heute als überholt betrachtet werden können, aber auch, dass sie spätere Wissensdarstellungsformalismen in einigen Grundannahmen beeinflusst

haben. Die Analogien zwischen Gedächtnis und Lexikon lassen sich auch bei moderneren Repräsentationsformalismen aus dem Umfeld der im Laufe der 70-er Jahre sich entwickelnden Cognitive Science feststellen. (Wissenschaftsgeschichtliche Aspekte der Cognitive Science sind in (Freska85, p.119ff) skizziert.) Bei Quillian ist diese Analogie offensichtlich und kommt in der Zusammenfassung von Wettler (p.40) gut zum Ausdruck:

"Quillian unterscheidet drei Arten von Verbindungen zwischen Knoten. Die erste Art von Verbindung assoziiert einen Begriff mit seinem Oberbegriff. Durch die zweite Art von Verbindung wird ein Begriff durch ein Adverb oder ein Adjektiv modifiziert. Die dritte Art von Verbindung dient dazu, zwei Wörter durch eine jeweils spezifizierte Relation miteinander in Beziehung zu setzen. Solche Relationen können Verben oder Präpositionen sein. Zusätzlich können in dem Formalismus von Quillian noch additive Relationen (Menschen haben Arme und Beine) und Disjunktionen (Nelken sind rot oder weiss) dargestellt werden."
(Wettler80, p.40)

In nuce sind in diesem Formalismus begriffslogische (Unterbegriff - Oberbegriff), grammatikalische (adverbiale oder adjektivische Modifikation), kasusgrammatische (Verben als Valenzträger, Präpositionen als Rektionsträger) aussagenlogische (Disjunktion) und mengentheoretische (additive Beziehung) Relationen eingebettet. Solche Relationen lassen sich sowohl in Wörterbüchern als auch in Wissensrepräsentationsformalismen der K.I. ("Künstliche Intelligenz") finden. Bei Wettler wird die K.I-Perspektive weiterverfolgt (Kasusgrammatik, Lexikalische Dekomposition, Frames), die Wörterbuchperspektive scheint weniger attraktiv zu sein.

Dass bei Quillians semantischem Gedächtnismodell (Netz) eine enge Beziehung zur Struktur eines Definitionswörterbuchs besteht, geht aus der einleitenden Bemerkung zum Arbeitsbericht

"Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities"
(Quillian67)

hervor:

"In order to discover design principles for a large memory that can enable it to serve as the base of knowledge underlying human-like language behaviour, experiments with a model memory are being performed. This model is built up within a computer by 'recoding' a body of information from an ordinary dictionary into a complex network of elements and associations interconnecting them. Then, the ability of a program to use the resulting model memory effectively for simulating human performance provides a test of its design."
(Quillian67, p.98)

Auch R.J. Brachmann weist explizit auf den Zusammenhang zwischen Quillians semantischem Netzformalismus und der Struktur eines Bedeutungswörterbuchs hin.

"In Quillian's structure, each word concept node was considered to be the head of a 'plane', which held its definition."
(Brachmann79, p.6).

Brachman bezieht sich in der obigem Zitat folgenden Stelle zu Quillian auf dieselben Relationen, die Wettler erwähnt:

subclass
modification
subject/object
disjunction
conjunction

Diese Relationen werden verwendet, um die Konzepte (Begriffe, Knoten) innerhalb eines "planes" (Definition) zu repräsentieren. Die Verbindungen zwischen den "planes" realisierte Quillian mit Verweisen (pointers) auf relationierte "planes" . Die pointer (Kantentypen, Relationen) sind limitiert. Durch die type-token Unterscheidung reduziert sich der Speicheraufwand.

"Each word was defined in only one plane in the structure (the head of the plane being the 'type' node), and all references to a word went through intermediate 'token' nodes. Thus definitions were not repeated each time a word concept was referenced."
(Brachman79, p.6)

Das graphisch dargestellte Beispiel einer solchen Konzeptdefinition und deren Verknüpfung mit anderen Konzepten, findet sich in (Quillian67) auf Seite 102. Dieses Beispiel der disambiguierten netzwerkartigen Darstellung des englischen Begriffs plant, das auch bei Brachman (p.7) und bei Wettler (p.41) angeführt wird, zeigt die Analogie zwischen Netzwerkformalismus und Bedeutungswörterbuch schlagend.

- PLANT. 1. Living structure which is not an animal, frequently with leaves, getting its food from air, water, earth.
2. Apparatus used for any process in industry.
3. Put (seed, plant, etc.) in earth for growth.

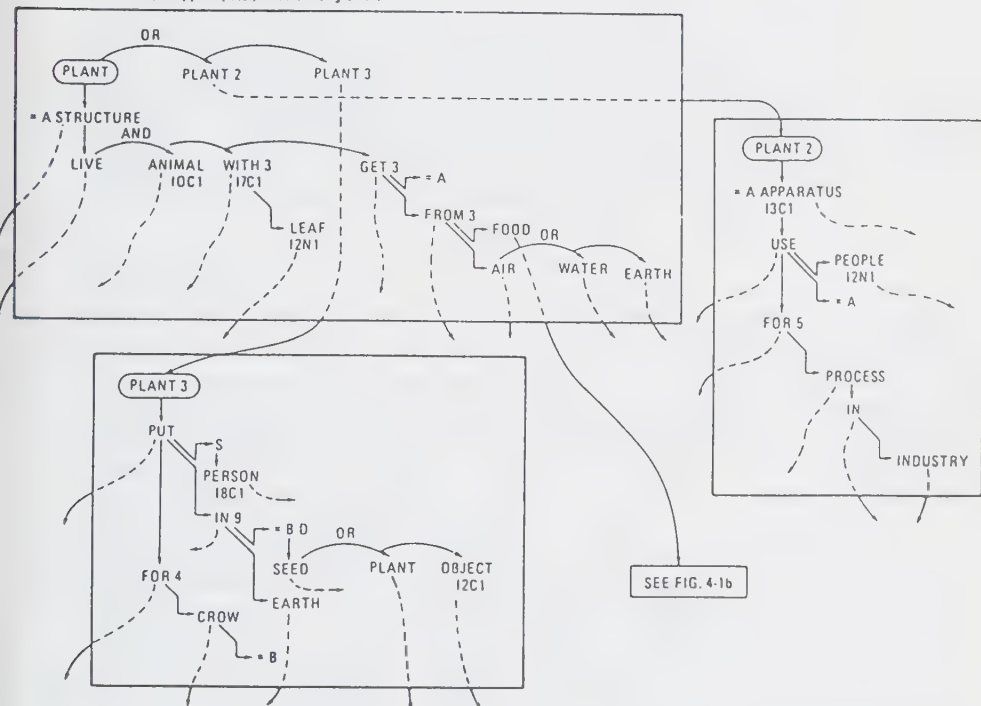


Figure 1. 'Frühes semantisches Netzwerk': nach (Quillian67)

Die Konsequenzen dieser Analogie sind besonders interessant, wenn man sich vor Augen hält, wozu der Netzwerkformalismus (das trifft auch auf andere Wissensrepräsentationen zu) dienen soll. Brachman stellt dazu lapidar und treffend fest:

"Quillian's desire of his semantic memory model was that it might serve as a general inferential representation of knowledge."
(Brachman79, p.6).

Und das ist genau das, was man von einer Wissensbasis erwartet. Die "naive" Interpretation von lexikographischen Abfolgescemata und deren syntaktische Strukturierung und Verknüpfung in Formen der

Wissensrepräsentation genügt aber nicht, um eine Wissensbasis für Inferenzprozesse - vorausagen ohne den Aufwand des Knowledge Engineering - aus einem Wörterbuch übernehmen zu können. Auch Quillian war nicht der Ansicht, mit relativ geringem Aufwand ein semantisches Gedächtnismodell aus einem besseren Bedeutungsörterbuch generieren zu können:

"Unfortunately, two years of work on this problem led to the conclusion that the task is much too difficult to execute at our present stage of knowledge"
(Quillian67, p.104)

Das war 1966 der Fall, aber heute (1986) kann man sich diese Mühe teilweise sparen und muss für die semantische Wissensrepräsentation nicht auf abgeleitete Wörterbuchstrukturen zurückgreifen, sondern auf elaborierte Repräsentationsformalismen aus der K.I. Es spielt dabei keine Rolle, ob solche Repräsentationsformalismen wie z.B. KL-ONE auf einfachere Netzrepräsentationen zurückgehen, die ursprünglich von lexikographischen Konzepten beeinflusst waren. Beispiele für dieses Vorgehen finden sich in den in Teil 2 dieser Arbeit vorgestellten Verfahren aus den Projekten HAM-ANS, KIT und EVAR.

Unabhängig davon, welchen Repräsentationsformalismus man benutzt, benötigt man vor der Realisierung einer Repräsentation eine genaue Analyse der zu repräsentierenden Wissensarten, Wissensquellen und Wissensprozesse.

Auf die Überlegungen bezüglich der formalen Komplexitätsstufen (a-f) von Wörterbüchern aufbauend, kann eine solche Wissens-Analyse skizziert werden. Wenn wir bei dem formalen Modell eines maschinenlesbaren Wörterbuchs mit den formalen Komponenten Kategorie (Element im Ablautschema), syntaktisch strukturiertes Ablaufschema (entspricht dem traditionellen Wörterbuchartikel) und der textgrammatisch strukturierten Makrostruktur bleiben, dann können alle diese Komponenten und ihre Verknüpfungen in Bezug auf Wissensarten, Wissensquellen und Wissensprozesse analysiert werden.

Bei einer derartigen Analyse muss nicht nur das im Wörterbuch zu repräsentierende Wissen berücksichtigt werden, sondern auch das beim Design implizierte Meta-Wissen. Bei der Unterscheidung zwischen lexikalischem Wissen und Metawissen handelt es sich um die Differenz zwischen dem real in den Einträgen codierten Wissen und dem Wissen, welches in den Anweisungen, wie diese Einträge zu generieren sind, vorausgesetzt wird. So kann ohne ein Meta-Wissen um die Wertebereiche der Kategorien kein einziges Abfolgeschema gebildet werden. Auch bei der Festlegung eines Codebooks oder einer Syntaktik für ein maschinenlesbares Wörterbuch werden Wissensquellen aktiviert, die im Wörterbuch selber nur noch zu einem geringen Teil expliziert sind, aber trotzdem für dessen Struktur konstitutiv sind. So sind z.B. strukturierte Inventare von semantischen Merkmalen oder das Repertoire von Kasusrahmen, Aktanten und Selektionsmerkmalen typische Meta Wissensquellen bezüglich des realisierten Wörterbuchs. Werden zwei Substantiven beispielsweise die semantischen Merkmale FLOWer oder INSTRUMENT zugewiesen, impliziert dies bei den meisten maschinenlesbaren Wörterbüchern ein Wissen über eine ausserhalb des Wörterbuchs festgelegte Organisation dieser Merkmale. (Bei den in dieser Arbeit untersuchten Systemen -vergl. Teil 2- trifft das z.T. für das Wörterbuch, das im Rahmen des Projekts KIT benutzt wird, nicht zu).

Zuletzt sind diese semantischen Merkmale eventuell im sog. Weltwissen verankert, aber für die Wörterbucharbeit liegt es in Form von klassifiziertem Merkmalwissen vor. Klassenbildung von Merkmalen, Ordnen dieser Klassen in Systeme und Zuweisung von Merkmalen zu den Wörtern sind Wissensprozesse, die der Erstellung des Wörterbuchs vorgeordnet sind.

Die Wissensanalyse erschöpft sich aber nicht in der Unterscheidung von Wissen und Metawissen: Nach dieser Differenzierung folgt eine Differenzierung des lexikalischen Wissens

- nach Wortarten,
- nach linguistischen Analysestufen (phonologisch, morphologisch, syntaktisch, semantisch, pragmatisch u.a.) und
- nach den vom verwendeten Repräsentationsformalismus verlangten Repräsentationslevels.

Zudem muss die Frage abgeklärt werden, ob die Repräsentation in einer einzigen Wissensbasis erfolgen soll oder ob mehrere Teilbasen notwendig sind.

1.1.2 Implementation.

Durch Hinweis auf Beispiele soll gezeigt werden, welche Implementationstechniken angewendet werden können. Diese Beispiele sind lediglich als Ergänzungen zu den oben erwähnten 7 Typen von Eintragsstrukturen (a-g) zu verstehen und nicht als Tool-Sammlung zum Thema Wörterbuch.

a) Zur Implementation von Wortlisten.

Die Implementation von Wortlisten muss nicht notwendigerweise mit der Datenstruktur Array oder Liste realisiert werden. Bei den in orthographischen Korrekturprogrammen für Textverarbeitung verwendeten speller-dictionaries werden z.B Kompressionstechniken verwendet, die es ermöglichen, auch umfangreiche Vokabulare auf Disketten zu speichern und schnell zu lesen.

"The word list is subjected to character frequency analysis, and the most common letters and combinations are replaced by shorter bit patterns. The program also puts the most common suffixes, likewise compressed, into a table. Within the table the most common groups of suffixes are reduced to small bit patterns. Several words can actually be represented with the same code. In an alphabetically sorted dictionary, most words have several letters in common with the previous word 'Lightning' takes advantage of this by storing only the word-to-word changes, not the common characters."

(Machrone85, p. 121)

Dieses Beispiel zeigt sehr schön, dass auch für eine logisch einfache Struktur, wie sie eine Liste darstellt, relativ aufwendige Implementationstechniken benutzt werden können, um die Effizienz der Implementation bezüglich Platzbedarf und Zugriffsgeschwindigkeit zu sichern.

b) Zur Implementation von zweiteiligen Abfolgeschemata.

Ein Beispiel für den bereits erwähnten Anwendungsfall für zweiteilige Abfolgeschemata in Lexikondateien für Parsersysteme ist das Vollformenlexikon für den Parser NEWCAT (Hausser85). Dieses Lexikon bestand 1985 aus 1000 Wortformen. Den Umfang dieser Datei gibt Hausser (p.12) mit

"90'000 Bytes (uncompiled) or 72'000 Bytes (compiled)"

an. Das Beispiel dieser Lexikodatei zeigt, dass auch ein zweiteiliges Abfolgeschema genügend Information enthalten kann, um den auf einer linksassoziativen Grammatik basierenden Parser NEWCAT mit Kategorien zu den Lesarten einer Wortform zu versehen. Ein Wortformeneintrag besteht in diesem Lexikon aus der Wortform und einer Kategorienliste.

Es könnte eingewendet werden, dass die Kategorienliste selbst aus mehreren Komponenten besteht und damit das Schema nicht mehr zweiteilig ist. Das ist auch richtig, aber andererseits hängt die formale Strukturbeschreibung vom Abstraktionsgrad der Betrachtung ab und die Liste kann abstrakt als ein Element aufgefasst werden.

Als Beispiel für einen Eintrag kann die Wortform "LAS" (von "lesen") angeführt werden:

```
<<LAS>
  <V S1 A>>
```

Für die Wortform "LAS" ist ein zweiter Eintrag erforderlich, weil diese Wortform morphologisch nicht eindeutig ist:

```
<<LAS>
  <V S3 A>>
```

In diesen beiden Einträgen ist die Oberflächenform (LAS) mit den Kategorien Verb (V), erste bzw. 3. Person Singular (S1 bzw. S3) und die Kasusangabe (A) für Transitivität codiert. Vom Tempus wird abstrahiert.

Für die Analyse werden die Einträge mit identischen Wortformen gesammelt und in eine grössere Einheit kombiniert:

```
[ <START(<<LAS>
  <V S1 A>>>
  <START(<<LAS>
    <V S3 A>>>)]
```

Das im so generierten Eintrag neue Element "START" ist ein Regelpaket, also Teil des Parsers NEWCAT. Der ursprünglich

zweiteilige Lexikoneintrag wird auf diese Art zu einer komplexeren Liste für die Analyse aufbereitet. Die Analyse skizziert Hausser so:

"Note that each reading of a word is analyzed as a sentence start, consisting of a rule package (START), and a list of the surface and the category of the word. Thus the combination of the first two words of a sentence or text is treated exactly like any subsequent combination: the parser forms the Cartesian product of the sentence start and the next word and applies the rule package listed as the first element of the first and second members of the pair."
(Hausser85, p.12)

*

Weitere Beispiele für die Implementation zweiteiliger Schemata können in bestimmten Techniken des Information Retrieval gefunden werden:

So kann für den in Kapitel 1.1.1 erwähnten hypothetisch einfachsten Fall von Information Retrieval aus Textdokumenten, (wo nur Dokumentbenennung und Dokumenttext unterschieden werden), ebenfalls von einem zweiteiligen Abfolgeschema gesprochen werden. Auch hier könnte eingewendet werden, dass besonders der zweite Teil des Schemas, der Text, so komplex sei, dass es nicht zulässig sei, von einem zweiteiligen Schema zu sprechen. Das ist aber nicht zutreffend, weil dieser Text nicht unter dem Aspekt seiner Komplexität betrachtet wird, sondern lediglich als eine Menge, über der mit treffenden Methoden Information Retrieval betrieben werden kann. Das hat dann zur Folge, dass diese Techniken subtil genug sein müssen, um dem Text ohne lexikonstrukturierende Eintragsmarkierungen die gesuchten Informationen entnehmen zu können. Nach U. Hahn können beim Retrieval über dem Text u.a. folgende Techniken angewendet werden: [**5]

1. Morphologische Verfahren:
Reduktion von Texttoken auf Grund- bzw. Stammformen.
vergl. (Hahn85b, p.55)
2. Syntaktische Verfahren:
Nominalgruppen-Erkennung, Dekompositionsverfahren, Partielles Parsing.
(Hahn85b, p.109)
3. Semantische Verfahren:
Statistische und inhaltsanalytische Methoden.
(Hahn85b, p.152)

Diese informationslinguistischen Techniken sind quasi der Ersatz für eintragstrukturierende Textmarken und werden benutzt, um die Texte abzusuchen und die gewünschten Teile zu finden.

*

Die beiden Beispiele zur Illustration des zweiteiligen Abfolgeschemas wurden mit Absicht so gewählt, dass zwischen ihnen nur diese formale "Ähnlichkeit" der Zweiteiligkeit besteht. Dies belegt einerseits die Dehnbarkeit der simplifizierenden lexikologischen Formel

zeigt aber andererseits, dass gerade auch derart "schwache" Formalismen in der Praxis tragfähig sind.

c) Zur Implementation von variablen Abfolgeschemata.

Hier kann zur Exemplifizierung auf konventionelle File und Recordorganisation verwiesen werden (VSAM, ISAM u.a.).

Ein besonders aufschlussreiches Beispiel für die Codierung variabler Abfolgeschemata bildet die Abbildung der morphosyntaktischen Angaben (MSA) im System der kumulierten Wortdatenbank ALEXSYS auf variable Recordstrukturen. Genauere Angaben zum System ALEXSYS sind in Kapitel 2.2.5 zu finden. Es handelt sich bei dieser Implementation um ein relativ komplexes Beispiel für variable Abfolgeschemata, weil morphosyntaktische Angaben aus verschiedenen schon bestehenden maschinenlesbaren Lexika kumuliert wurden und die Angaben in den Originallexika erhalten bleiben sollen. Die Datenstruktur für die MSA in ALEXSYS ist auch deshalb komplizierter als die in der formalen Beschreibung (1.1.1(c)) entwickelte lineare Struktur des variablen Abfolgeschemas, weil zur Eigenschaft der variablen Länge noch die Hierarchie hinzukommt.

Innerhalb der einzelnen Hierarchieebenen lässt sich aber an diesem Beispiel sehr gut demonstrieren, wie durch repetitive Gruppen variable Abfolgeschemata implementiert werden. Die hierarchischen Aspekte dieses Beispiels bilden zugleich eine Illustration für die baumstrukturierten Einträge, auf die dann aber in (1.1.2(e)) dieses Kapitels nochmals eingegangen wird. Das Beispiel zeigt auch, dass in der Praxis die formal entwickelten Typen selten rein vorkommen. Die folgende Darstellung basiert auf (Schulze83):

"Die ersten 40 Bytes sind mit der Wortform belegt. Es folgt eine aus 2 Bytes bestehende Bit-Leiste, in deren ersten 12 Bits die Information verschlüsselt ist, in welchem der kumulierten 12 Einzellexika ein Eintrag zu der Wortform zu finden ist (Belegstellen-Angabe oder Lexikon-Kennung).

Zwei weitere Bits enthalten eine grobe Grund-/Vollformenmarkierung, d.h. sie geben Aufschluss darüber, ob eine Wortform als Grund-Vollform zu interpretieren ist [**6].

Das folgende 20-Bytes-Feld enthält alle WK-Codes [Wortklassen H.H.] zu der betrachteten Wortform, unabhängig vom jeweils bearbeiteten Einzelfile [**7], sowie Markierungen für die interne Verarbeitung.

Für reine Grundformen ist keine Angabe eines Lemmanamens vorgesehen [**8], sodass hier die morphosyntaktischen Angaben, beginnend mit '&', bereits auf Position 63 anfangen.

Für Vollformen hingegen findet sich auf Position 63 der zugehörige Lemmaname [**9]."

(Schulze83, p. 35)

Nun beginnen die Codierungen der eigentlichen morphosyntaktischen Angaben, die nicht nur Wiederholungsgruppen aufweisen, sondern auch

variable Codebedeutungen in Abhängigkeit von vorangehenden Codierungen. Auch diese Eigenschaft der Datenstruktur geht über den linearen Charakter der formal skizzierten variablen Abfolgeschemata hinaus. Hinzu kommt, dass das System ALEXSYS in vier grosse Files gesplittet ist, in der die hier geschilderte Datenstruktur unterschiedlich interpretiert wird. (Diese 4 Files sind: 1. Nomen und Namen, 2. Verben, 3. Qualifier, 4. Funktionswörter).

"Mit jedem "&"-Zeichen beginnt in der Datenstruktur ein neues Gruppenfeld, das einen Wortklassen-Code (WKC) und die von diesem WKC abhängigen morphosyntaktischen Angaben enthält. Dabei werden alle die WKC berücksichtigt, die im jeweiligen File relevant und zu der gegebenen Wortform gesetzt sind."
(Schulze83, p.35)

Innerhalb dieser Wiederholgruppe kommt wieder eine variable Gruppe von Feldern, die die MSA enthalten. Zu Beginn der Gruppe steht die Angabe, aus wievielen Feldern sie besteht, dann kommen die Felder. Diese Felder bestehen jeweils aus 3 Bytes. Und dabei

" sind die einzelnen MSA-Felder in ein 2-Byte-Feld und in ein 1-Byte-Feld unterteilt:

Das 1-Byte-Feld (im Schaubild TP1, FM1) enthält im VRB-File Angaben über das 'transformatorische Potential', im FWK-File über die 'transformatorische Initiative', im NOM/NAM-File und im QAL-File flexionsmorphologische Angaben, die sowohl für Grund- als auch für Vollformen relevant sind. Dieses Feld ist stets als Bit-Leiste zu lesen, wobei das erste Feld darüber Auskunft gibt, ob es sich um ein MSA-Feld zu einer Grundform (Ausprägung '0') oder Vollform (Ausprägung '1') der gegebenen Wortform handelt."
(Schulze83, p.35)

Dabei handelt es sich also um das 1. Bit des 3. Bytes der ganzen Dreiergruppe oder anders gezählt um das 17. Bit. Dieses Bit 17 wirkt wie ein Interpretationsschalter auf die beiden vorangehenden Bytes, die je nach Switchstellung unterschiedlich zu interpretieren sind:

"- Für Vollformen ist es [das 2-Byte-Feld H.H.] als Bit-Leiste zu lesen, in der die nur für Vollformen relevanten flexionsmorphologischen Angaben verschlüsselt sind ('ist'-Zustand der Wortform)

- Für Grundformen ist es als Charakter-String zu lesen und enthält einen Code für die zutreffende Flexionsklasse ('kann'-Zustand der Wortform)."
(Schulze83, p.35)

Die folgende Abbildung aus (Schulze83, p.36) zeigt die ganze Datenstruktur.

Datenstruktur in den Files der kumulierten MSA

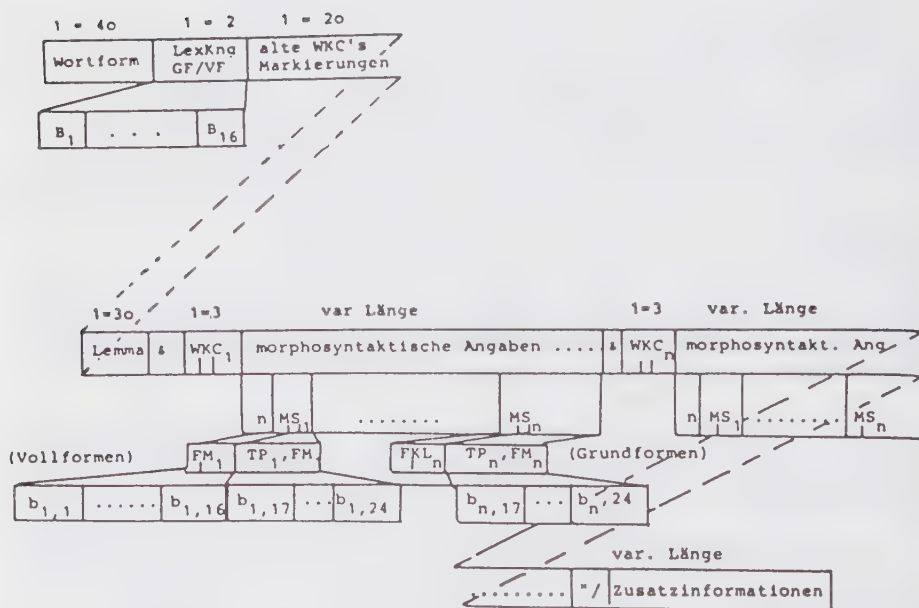


Figure 2. 'Morphosyntaktische Angaben (ALEXSYS)'

Die Möglichkeit, variable Abfolgeschemata mit Recordstrukturen zu implementieren, *wird heute durch andere Möglichkeiten, die flexibler und transparenter sind, ergänzt:

Anstatt Variationen durch implizites Displacement abzubilden, kann die Technik des Tagging verwendet werden. Dabei werden wie in einer Mark-up-language die Kategorien vor ihrem Auftreten - und nur bei diesem - mit einem Tag signalisiert, womit das nachfolgende Element als Mitglied der durch das Tag markierten Kategorie erkennbar wird.

"Jedes sprachliche Objekt, nicht nur der Wörterbuchtext, dessen Textteile auf die geschilderte Art etikettiert werden können, nennen wir hernach einen markierten Text Kurz und englisch gesprochen unterscheiden wir 'tag' und 'text', Etikette und Textstück.

Ein Wörterbucheintrag lässt sich nun verhältnismässig einfach als eine geordnete Menge von tag-&-text-Elementen vorstellen" (Drewek85, p36)

d) Zur Implementation von Tabellen.

Ein Beispiel für die Implementation von Wörterbüchern mittels Tabellen stellt der Datentyp "Wörterbuch" in der Programmiersprache COMSKEE dar, welche in Saarbrücken (sfbl100) entwickelt wurde. "Wörterbuch" ist ein dynamischer Datentyp und kann im Prinzip als

"Wörterbuch-Variable" "genauso benutzt" werden, "wie gewöhnliche array-Variablen" (Messerschmidt84, p.150).

Die Wörterbuchvariable ist dabei flexibler als der in COMSKEE ebenfalls definierte Datentyp Array, weil

"auf der Indexposition beliebige Ausdrücke, die einen beliebigen (string-, number-, oder bits-) Wert berechnen, stehen können", (Messerschmitt, p.150)

und weil Wörterbuchvariable auch Eigenschaften vom Datentyp File haben (p.151). Interessant ist bei dieser Konzeption, dass der Datentyp "Wörterbuch" vollständig in eine Programmiersprache integriert ist, in der u.a. auch ein Datentyp "sentence" definiert ist. Dieses Typenkonzept erinnert an die Datentypen in Pascal, nur sind in COMSKEE diese Datentypen auf linguistische Bedürfnisse ausgerichtet. Insofern können sie als spezifischere Varianten der abstrakten Datentypen (ADT) im Sinne von (Aho83) aufgefasst werden.

Ist die formale Struktur des Wörterbuchs als Tabelle oder als Tabellenkombination konzipiert, bietet sich für die Implementation das Modell relationaler Datenbanken an, welches auf Codd [**10] zurückgeht. Vom Anwenderstandpunkt her betrachtet, ist dieses Modell sowohl in (Zehnder85, p.34-77), als auch in (Vetter82, p93ff) besonders im Hinblick auf "Normalisierung" beschrieben. Durch den Normalisierungsprozess können die Tabellen (Relationen) in mehreren Stufen in eine "Normalform" transformiert werden. In einer 3. Normalform wird eine redundanzarme Speicherung erreicht und

Änderungen in den Tabellen führen in der Datenbank relativ selten zu Anomalien. [**11] Mehrfachspeicherungen sind auf die Schlüssel beschränkt. Es ist für grosse Datenmengen, wie sie Lexika darstellen, jedoch nicht unbedingt optimal, bis zur 3. Normalform zu normalisieren, weil sonst die Information in zu vielen Tabellen verstreut wird und Zugriffe dadurch zu aufwendig für das DBMS werden. Allerdings kann in relationalen DBM-Systemen durch Clusterbildung diesem Mangel abgeholfen werden.

Falls ein relationales DBMS angewendet werden soll, gliedern sich die Arbeiten (stark vereinfacht betrachtet) in:

1. Analytische Vorarbeiten, um zur Festlegung eines lexikalischen Datenmodells zu gelangen. (Eine Analyse ist in jedem Fall notwendig, unabhängig davon, ob ein relationales DBMS verwendet werden soll).
2. Normalisierungsprozesse.
3. Entwurf und Implementation.

Das Datenmodell schliesst sowohl die Vereinbarung der Datenstruktur, als auch die Vereinbarung der auf der Struktur möglichen und zugelassenen Operationen ein. In der Entwurfs- und Implementationsphase stellt die Datenbanktechnik dafür Sprachen zur Verfügung, mit deren Hilfe die Datenstruktur festgelegt wird und die Operationen programmiert werden können.

Bevor an eine Programmierung des Datenmodells gedacht werden kann, muss das Modell vor allem von der linguistischen Applikationslogik her ausformuliert werden. Für die Datenanalyse eignet sich dabei das Entity-Relationship-Model, welches auf Chen [**12] zurückgeht. vergl. (Zehnder85, p.34ff). Weitere Modelle sind in (Zehnder85, p.94) erwähnt.

Wenn dieses Modell nicht auf Maschinenteile, Stücklisten, Projekte, Mitarbeiter oder Bankkonten angewendet werden soll, sondern auf lexikalische Entitäten und Relationen, bedingt dies eine in bezug auf eine bestimmte Datenbanktradition eher ungewöhnliche Analyse des zu modellierenden Lexikons und der mit diesem Lexikon geplanten Anwendungen. [**13] Besonders ist auf Integrität der Daten zu achten, welche nach Zehnder Konsistenz, Datensicherheit und Datenschutz einschliesst. Integrität kann durch die Beschränkung der zugelassenen Operationen auf konsistenzerhaltende Operationen erzwungen werden.

Einen allgemeinen Überblick über "Integritätsaspekte in Datenbanksystemen" geben D. Steinbauer und H. Wedekind (Steinbauer85, p60). Vergl. zu Konsistenz auch den nächsten Abschnitt (e. Baumstrukturen), wo am Beispiel der lexikographischen Datenbank HBASE des Systems EVAR auf die Vielfalt spezifisch lexikographischer Datenbankoperationen hingewiesen wird.

In relationalen DBMS wird durch die in der Abfragesprache formulierbaren Kombinationen von Selektion, Projektion und Join die zu extrahierende Informationsmenge in der Regel nach beliebigen Suchkriterien zugänglich, nicht nur über das Lemma, wie in einem gedruckten Wörterbuch.

An dieser Stelle kann noch eine Anmerkung zur Verwendung des Begriffs "Datenbank" im Zusammenhang mit Wörterbüchern gemacht werden. Die in diesem Zusammenhang manchmal verwendeten Begriffe "Sprachdatenbank", "Worddatenbank" und "lexikalische Datenbank" sind erst klar, wenn angegeben wird, ob es sich bei einer solchen Datenbank um

- das zugrundeliegende DBMS als Implementationstool,
- um eine Sammlung mehrerer maschinenlesbarer Lexika,
- oder lediglich um ein maschinenlesbares Wörterbuch in irgend einer Form handelt.

In der Praxis hat sich allerdings ein relativ allgemeiner Gebrauch des Begriffs "Worddatenbank" eingestellt:

"Die Bezeichnung 'Worddatenbank' für ein Anwendungsgebiet maschinenlesbarer Wörterbücher ist noch relativ jung, obwohl Worddatenbanken der Sache nach beinahe seit Beginn der linguistischen Datenverarbeitung existieren. der Terminus wird zur Bezeichnung einer jeden Sammlung lexikalischer Einheiten einer Sprache, versehen mit einer Menge von Informationskomplexen, verwendet, auf die im Rahmen eines Datenbanksystems zugegriffen werden kann und die durch ein Datenbanksystem verwaltet werden. Im allgemeinen beruhen diese Sammlungen auf Auswertungen von Korpora, bereits gedruckten oder auch maschinenlesbar vorliegenden Lexika. Zu nennen wären hier Worddatenbanken, die - zum Teil unter der Bezeichnung 'Terminologiedatenbanken' - in Kanada, Frankreich, Grossbritannien (Sager/MC Naught 1980) und Belgien in Angriff genommen wurden. Für das Deutsche liegen Pläne bzw. Vorarbeiten [jetzt realisiert, H.H.] für eine lexikographische Datenbank (IDS, Mannheim) und für eine kumulierte Worddatenbank maschinenlesbarer Lexika (IKP, Bonn) vor." (Hess83, p22/23)

Als "Vorläufersysteme" bezeichnen Hess/Brustkern/Lenders

"die maschinenlesbare Version des Wörterbuchs von Lutz MACKENSEN und die sogenannte WAHRIG-Datenbank" (Hess83, p23)

(Vergl. dazu Gudrun Guckler, Evelyn Müller, Gerhard Wahrig, Wörterbuch als Datenbank. Ein komplexes linguistisches, logisches und technisches Problem. In: (Wahrig83), p. 350-440).

e) Zur Implementation von Baumstrukturen.

Zur Illustration der Implementation von baumstrukturierten (hierarchischen) Lexikoneinträgen eignet sich das in Teil 2 (Kapitel 2.2.3 und 3.3.2) dargestellte Lexikonsystem EVAR. In diesem System ist ein Worteintrag baumartig strukturiert, nämlich:

"in einen phonologischen Teil, einen Syntaxteil, der mehrere Worteinträge beinhaltet, jeder einzelne Wortarteintrag einen Semantikeil enthaltend, der wiederum mehrere Bedeutungen mit eventuell jeweils mehreren Aktanten umfasst. Aufgrund dieser Struktur war es auch das Ziel, durch einen entsprechend gegliederten, hierarchisch strukturierten Programmablauf die Möglichkeit zu schaffen, auch einzelne Teile dieses Wortbaumes - Syntaxteil, Semantikeil oder Aktanten - als kleinere Einheiten ebenfalls portieren zu können. (Endres86, p.50)

"Unter Portierung wird in dieser Arbeit die Umsetzung der Textdateien des Lexikons in die hier vorgestellte interne Datenstruktur verstanden." (Endres86, p.50)

Die Implementierung des Datenmodells besteht

1. in der Datenstruktur des Lexikons mittels des für EVAR modifizierten Datenbanksystems HBASE (nähere Angaben zu HBASE siehe Kapitel 2.2.3) und
2. in den über HBASE in Lisp definierten lexikographischen Funktionen.

Die nun folgende Auflistung der in Franz-Lisp programmierten lexikographischen Funktionen basiert auf (Endres86):

a) Änderungsfunktionen:

- aend-aussprache
- aend-flekt-formen
- aend-syneintrag
- aend-semeintrag
- aend-akteintrag

b) Portierungsfunktionen:

Diese betreffen die Umsetzung in die interne Datenstruktur. Bei dieser Umsetzung werden gleichzeitig Konsistenztests durchgeführt.

c) Konsistenztests:

Endres weist darauf hin, dass es sehr aufwendig ist, inhaltliche Konsistenzchecks zu programmieren. Formale Konsistenztests sind:

- kons-syneintrag
- kons-semeintrag
- kons-akteintrag

Inhaltliche Konsistenztests sind lediglich als Dummy-Funktionen eingebaut (syn-konsistenz, sem-konsistenz, akt-konsistenz) und später ergänzt. (p.55) (Stand 04/86).

d) Auskunftsfunktionen:

Die Auskunftsfunktionen sind als Benutzerschnittstelle konzipiert und heissen:

- wortp. Diese Funktion ermöglicht die Wortsuche in verschiedenen Formaten, u.a. über die Wortnummer.
- wortartp. Liefert zu einem Wort die Wortart.
- bedeutungp. Liefert zu einem Wort die Bedeutung
- akt-realp. Diese Funktion liefert die Aktantenrealisierungen zu gegebenem Syntaxitem.

Mit get-akt-spez-syn, get-aktspez-sem und get-akt-spez kann zusätzlich angegeben werden, ob nach fakultativen oder obligatorischen Aktanten gesucht wird.

Eine ganze Reihe weiterer Auskunftsfunktionen spezifizieren jeweils eine grössere Menge von die Suchkriterien erfüllenden Returnwerten, beispielsweise die Liste aller Semantikitems zum gegebenen Syntaxitem. (p.60-62)

e) Suchfunktionen:

Die Suchfunktionen sind komplexer als die Auskunftsfunktionen und werden durch ein Such- bzw. Anforderungsprofil bestimmt. Nach Aufbereitung des Suchprofils wird die Funktion suche-lex-sem-akt-syn

aktiviert. Diese Suchstrategie ist auf p.66 beschrieben.

f) Generierung eines Stammllexikons:

Diese Funktion ergibt sich daraus, dass das Lexikon in EVAR ein Vollformenlexikon ist. (p.71ff)

g) Ausgabefunktionen:

Die Ausgabefunktionen ermöglichen ein variables Layout des Lexikons (p.78)

h) Hilfsfunktionen:

Dazu zählt z.B. eine trace-Funktion, welche die Abfolge der Funktionsaufrufe protokolliert. Weitere Hilfsfunktionen sind bei (Endres86) auf den Seiten 81-87 beschrieben.

*

Wenn man die Implementation von Baumstrukturen nicht von der generativen Seite her betrachtet (d.h. Erzeugung von hierarchischen Lexikoneinträgen), sondern von der analytischen Seite her (d.h. Parsen von Wörterbüchern), kann als Beispiel für eine Implementation auf NET/1 (Drewek84) hingewiesen werden:

"NET/1 leistet beim Abruf von Einträgen aus der Datenbasis ein Parsing, dessen Ergebnis die Baumstruktur des Eintrags ist. Das in der Hierarchie gespeicherte Metawissen ermöglicht erst intelligente Operationen wie z.B. die korrekte Umstülpung von Quell- und Zielsprache, das Aufbrechen von Nestern oder die korrekte Auflösung von Tilden. Aber auch der Inhalt der Textstücke eines Eintrags kann mit allen gängigen Textbearbeitungsfunktionen verändert werden." (Drewek85, p.42)

f) Zur Implementation von Wörterbuchsyntaxiken.

Ein Hilfsmittel zur Definition einer Wörterbuchsyntax ist die Verwendung eines automatischen Sprachübersetzers (Compiler).

Bei einem Compiler besteht der Input in einem highlevel-Code für ein Programm oder in einer Data-Definition für ein Schema einer Datenbank und der Output in der maschinennäheren Form des Inputs. Wie viele Schritte bei diesem Compilationsvorgang durch ein Entwurfssystem (oder Präprozessor) automatisiert sind, ändert nichts am Charakter dieses Übersetzungsprozesses. (Vergl. Zehnder85, p.69ff., Der Weg zum konzeptionellen Schema mit Hilfe des Entwurfssystems Gambit). Der Compiler übersetzt dabei den Input nach den syntaktischen und semantischen Regeln der Syntax der high-level Sprache in die maschinennähere Version des Outputs. Im Falle von Gambit bildet das Entitätenblockdiagramm, ergänzt um die Attributsangaben, die Wertebereiche und die Konsistenzbedingungen den Input. Neben dieser interaktiven Unterstützung beim Entwurf des Entitätenblockdiagramms (Zehnder85, p.69), besteht die Leistung des Entwurfssystems Gambit in einer "Transformation des ergänzten konzeptionellen Schemas in maschinenverwendbare Form (Codierung)" (Zehnder85, p.69). Der Output bildet die "Zielsprache, die 'maschinenverwendbare Form' eine eigentliche Datenbanksprache (Modula/R)" (Zehnder85, p.69). Modula/R ist ein in Modula-2 eingebettete relationale Datenbanksprache im Sinne von Codd. (Zehnder85, p.77-82). Zusammenfassend formuliert Zehnder die drei

Aufgaben von Gambit wie folgt:

1. "Interaktive Unterstützung beim Entwurf des Entitäten-blockdiagramms"
2. "Gezielte Ergänzungen des konzeptionellen Schemas"
3. "Transformation des ergänzten konzeptionellen Schemas in maschinenverwendbare Form (Codierung)"

Die Vorteile eines solchen Systems werden von Zehnder hervorgehoben:

"All diese Aufgaben müssen auch erledigt werden, wenn eine Datenbank aufzubauen ist, ohne dass ein Entwurfssystem zur Verfügung steht. Während in einem solchen Fall die ersten beiden Schritte, nämlich Entwurf und Ergänzung des Entitätenblockdiagramms, ausschliesslich manuell mit Papier und Bleistift (und Gummy!) ausgeführt werden müssen, wobei sich immer wieder Fehler einschleichen können, kann der Schritt der Codierung je nach verwendeter Sprache und Compiler maschinell unterstützt werden."

(Zehnder85, p.70)

Bezogen auf Wörterbücher bedeutet Compilation in der Regel, dass ein Übersetzer geschaffen wird, der die Syntax "versteht", nach welcher das Wörterbuch aufgebaut ist. Dieser Compiler hat die Fähigkeit, Einträge zu erfassen und in eine Worddatenbank abzulagen. Benutzerfreundlich ist ein solches System, wenn die Erfassung der Einträge vor der Compilation in nach linguistischen Kriterien aufgebauten Masken und Fenstern erfolgen kann. Dazu würde auch kontextabhängige Helpinformation gehören, welche für jede Kategorie eines jeden Eintrags die zur Auswahl stehenden Codierungsmöglichkeiten in aufgeschlüsselter Form auf Wunsch anzeigt.

Aus der Sicht des Systemspezialisten bildet die Verwendung eines Compiler-Compilers ein Hilfsmittel zum Erstellen eines Wörterbuch-Syntax-Compilers. Der Compiler-Compiler erhält als Input die Syntax des Wörterbuchs und generiert einen Compiler, welcher Einträge ins Wörterbuch "versteht" und in eine Datenbank ablegen kann. Dieses gestufte Vorgehen ermöglicht Experimente mit variabler Syntax: Anstatt für jede veränderte Wörterbuchsyntax einen neuen Compiler schreiben zu müssen, genügt es, dem Compiler-Compiler die neue Syntax einzugeben, worauf ein entsprechender Compiler generiert wird, welcher dann (ebenso wie der "von Hand" geschriebene Compiler) in der Lage ist, die Syntax der in den generierten Compiler eingegebenen Wörterbucheinträge zu "verstehen".

Erweitert man dieses Konzept und wendet es auf maschinenlesbare Wörterbücher an, wie das (Domenig86) tut, könnte ein derart gestuftes Vorgehen, von der Anwenderseite her betrachtet, so benutzt werden, dass der linguistische Spezialist in der ersten Stufe das System mit hauptsächlich grammatikalischen Regeln (Morphologie: Flexionsregeln, Derivationsregeln, Kompositionsregeln u.a.) versieht, und dass dann das derart gerüstete System einen programmierten Dialog mit dem lexikographischen Spezialisten generiert, in welchem die für das Wörterbuch erforderliche lexikalische Information akquiriert wird.

Der Nutzen einer solchen Konzeption besteht u.a. in der gesteigerten Variabilität bezüglich Regeln und Sprachen. Es können

sinnvollerweise nicht nur eine Lexikon-Syntax generiert werden, sondern mehrere, je nach Art des gewünschten Lexikonsystems, bzw. der in der Wortdatenbank berücksichtigten lexikalischen Regelmässigkeiten.

Ein einfaches Beispiel für das Funktionieren eines Compiler-Compilers bildet "Lex", ein Scanner-Generator, der unter dem Betriebssystem UNIX zur Verfügung steht. (M.E. Lesk, E. Schmidt, Lex: A Lexical Analyzer Generator, Bell Labs, Murray Hill, 1980). Der vollständige Compiler-Compiler Yacc kann Lex kontrollieren. (St. C. Johnson, Yacc: Yet Another Compiler-Compiler, Bell Labs, Murray Hill, 1981). Lex funktioniert aber auch ohne Yacc. Zur Illustration der grundsätzlichen Funktionsweise eines Compiler-Compilers, genügt es in diesem Fall, nur Lex zu betrachten. Lex kann beispielsweise dazu verwendet werden, aufgrund der Regeln der einzugebenden Syntax einen Scanner zu generieren, welcher einen Text "verstehen" (abtasten) soll. In einem zweiten Schritt wird der so generierte Scanner dazu verwendet, den Text tatsächlich zu "verstehen". Um ein sehr einfaches Beispiel zu nehmen, könnte man einen Scanner generieren, welcher Stopwörter in einem Text unterdrückt, aber die übrigen Wörter des Texts ausgibt. Die Syntax, nach der dieser Scanner konstruiert werden soll, wird "Lex" als Input übergeben und besteht aus der Definition der Wortseparatoren, der Angabe der als Wörter erlaubten Zeichenketten sowie einer Liste von Stopwörtern, die, falls sie im Text vorkommen, nicht ausgegeben werden dürfen. Aufgrund dieser Syntax erzeugt Lex einen Scanner, der später für die eben definierte Aufgabe der Stopwortelimination eingesetzt werden kann. [**14]

g) Zur Implementation von lexikalischen Wissensbasen.

Es gibt keinen Grund, warum Wissensrepräsentations - Formalismen aus der K.I. für Wörterbücher nicht angewendet werden können. Wörterbuchstrukturen können als Pattern directed Modules einer Wissensbasis angesehen werden, die auf definierte Match-Bedingungen sensitiv sind. Das kann z.B. eine in PROLOG geschriebene Regel- und Faktenbasis sein, in welcher lexikalisches Wissen enthalten ist.

Eine Alternative zur direkten Abbildung des Wörterbuchwissens auf eine Knowledge Representation Language ist die Möglichkeit, das im Lexikon enthaltene Wissen über einen Zwischenprozess in die für die Wissensrepräsentationskomponente des Systems kompatible Form zu transformieren, wie das z.B. im System EVAR durch das Nulllexikonkonzept gemacht wird. (Vergl. Kapitel 2.2.3).

Zwei Beispiele für die direkte Repräsentation des Lexikonwissens in einem Wissensrepräsentationsformalismus sind beschrieben in:

1. I. Wegner, Frame Theorie in der Lexikographie (Wegner85) und

2. L. Rostek, D. Fischer, Konzept für ein thesaurusbasiertes Information Retrieval am Arbeitsplatz. (Rostek85a)

I. Wegner modifiziert Frame Repräsentationen von Winograd, Minsky, Lehnert, v. Dijk und Ballmer, wobei er besonders auf (Ballmer77) hinweist:

"Bei meinem Vorgehen berücksichtige ich Ballmers sprachliches Relevanzprinzip, nach dem der Wortschatz in Gestalt höherstufiger Lexeme Möglichkeiten zu seiner eigenen Strukturierung enthält, indem ich Framedimensionen verwende. Die Frames stellen in Ballmers Ausdrucksweise richtige Parametrisierungen semantischer Strukturen dar. (Wegner85, p.55)

An 5000 Substantiven aus dem Lexikon von (Dornseiff33) demonstriert Wegner, dass ein

"explizites onomasiologisches Wörterbuch" ,
"dessen Einträge frameähnliche Struktur aufweisen",
(Wegner85, p.46)

realisierbar ist. Er beschränkt sich dabei auf einen Teilbereich von Substantiven, nämlich die Wortgruppe "Gesellschaft und Gemeinschaft". (Das zum Teil veraltete Vokabular und die ebenfalls z.T. fragwürdige Sachgruppenteilung im Dornseiff wird gegenwärtig neu bearbeitet, vergl. Kaiser84)

*

L. Rostek (Rostek,85a,85b auch: Rostek82) entwickelt eine semantische Begriffsrepräsentation, die sich wie die Frame Repräsentation Wegners auf den nominalen Bereich bezieht. Es geht dabei um die Information Retrieval Komponente eines Autorenarbeitsplatzsystems. Die Speicherung und das Suchen von durch beliebige Nominalgruppen benannten Dokumenten funktioniert über ein dreistufiges Repräsentationsmodell der zu verwaltenden Dokumentbenennungen.

Die 3 Stufen werden mit Modell 0, Modell 1 und Modell 2 bezeichnet:

- Modell 0 ist die morphologisch normalisierte Deskriptorenmenge,
 - Modell 1 erkennt dank syntaktischem Parsing Bindungen zwischen Deskriptoren und in
 - Modell 2
- "schliesslich wird durch semantisches Parsing ein Ausdruck in einer Prädikatenlogik 1. Stufe mit Typen oder 'roles' oder Tiefenkasus für die Argumentstellen generiert"
(Rostek85a, p.10).

Das Ziel der Repräsentation umschreibt Rostek wie folgt:

"Ein solches Instrument hat insbesondere nicht-eindeutige Benennungen zu verwalten - eine Aufgabe, die eigentlich von jedem kooperativ arbeitenden Computersystem zu leisten wäre."
(Rostek85a, p.9)

Die Aufgabe der semantischen Repräsentation von Nominalgruppen durch "Modell 2" besteht also darin, die vom Benutzer vergebenen Benennungen mit semantisch ähnlichen zu assoziieren, um so eine dem Individuum angepasste Mehrfachklassifizierung von Dokumenten und ein benutzerfreundliches Retrieval zu ermöglichen.

*

Diese beiden Beispiele sollen die Plausibilität der Perspektive "Lexikon als Wissensbasis" im Hinblick auf Praxis und

Implementation belegen. Weitere Beispiele bilden die in dieser Arbeit untersuchten Wörterbücher der drei KI-Projekte HAM-ANS, KIT und EVAR (vergl. Teil 2 dieser Arbeit).

Bei der Realisierung des Konzepts "Lexikon als Wissensbasis" sind aber eine ganze Reihe von nicht zu unterschätzenden Schwierigkeiten zu berücksichtigen. In einem Protokoll-Ausschnitt zum Arbeitstreffen [**15]

"Spezifikation einer lexikalischen Wissensbasis (Deutsch) für natürlichsprachliche KI-Systeme"

wird in kompakter Form auf diese immer noch bestehenden Schwierigkeiten hingewiesen. Es ist zu beachten, dass diese Schwierigkeiten nicht technischer, sondern konzeptioneller Art sind. Der Protokoll-Ausschnitt ist auch deshalb von Interesse, weil an diesem Treffen u.a. Vertreter der drei Projekte referierten, welche im empirischen Teil dieser Arbeit vorgestellt werden. (M. Eimermacher, Anforderungen des KIT Projektes an ein Lexikon; H.-W. Hein, Die linguistische Datenbasis des Erlanger Spracherkennungssystems; W. Hoepfner, Lexikalische Wissensquellen im Dialogsystem HAM-ANS).

"Folgende wesentlichen Aussagen und Fragen (sie sind als Meinungsäußerungen und nicht als Meinung des Plenums zu verstehen) wurden protokolliert:

- * Die existierenden grossen Computer-Lexika sind für KI-Systeme weniger geeignet.

- * Welches sind die kleinsten Informationseinheiten in KI-Lexika?

- * Wie soll man Interessantheitsgrade und Evidenzfaktoren in ein KI-Lexikon integrieren?

- * Die Semantik ist im allgemeinen schwierig abzuhandeln, weil sie projektbezogen ist.

- * Man sollte sich über Wortklassensysteme einigen.

- * Wie findet man inhaltlich abgrenzbare Wortmengen? Durch die Analyse von Texten aus dem entsprechenden Gegenstandsbereich?

- * Eine lexikalische Wissensbasis sollte von KI-Projekten als 'Steinbruch' benutzt werden.

- * Die Semantik ist das wesentliche Problem bei natürlichsprachlichen KI-Systemen.

- * man sollte das Arbeitsthema nicht zu sehr im Bezug zu laufenden Projekten sehen, sondern für zukünftige Projekte planen.

- * Kann man semantische Information einem Lexikon überhaupt inkrementell beifügen, z.B. Verb für Verb? Oder ist das etwas, das über den einzelnen Worteintrag hinausgeht?

- * Linguistische Prozeduren in KI-Systemen sollten universell sein. Projektabhängige Semantik ist ein Witz.

* Eine umfassende Analyse deutscher KI-Systeme auf die tatsächlich enthaltenen semantischen Definitionen ist notwendig.

* Ist Semantik empirisch fundierbar, oder muss sie immer linguistisch fundiert sein?

* Zum Thema Semantik sollte man sich an Vorhandenem orientieren, da dies offensichtlich formulierbar ist.

* Gibt es generell einen Grundwortschatz?

* Ist ein gemeinsamer Wortschatz für den Leistungsvergleich bei natürlichsprachlichen KI-Systemen notwendig?"
(H.-W. Hein83, p.28ff)

Diese informelle Auflistung von Meinungen von Experten zeigt immerhin, wo etwa die Schwierigkeiten liegen könnten, die sich bei der Implementation lexikalischer Wissensbasen mit semantischer Information ergeben.

1.1.3 Adäquatheit der Repräsentation.

Die in 1.1.1 entwickelten Typen von Wörterbüchern und die in 1.1.2 erwähnten Implementationstechniken sind dann als Repräsentation von lexikalischem Wissen geeignet, wenn sie den nun folgenden Adäquatheitskriterien genügen. Die Adäquatheit der Repräsentation des lexikalischen Wissens sollte

- formal,
- wissenstheoretisch und
- linguistisch

gegeben sein. Diese drei Termini werden im folgenden expliziert (a-c).

a) Formale Adäquatheit.

Die formale Adäquatheit der Repräsentation der Eintragsstruktur eines maschinenlesbaren Wörterbuchs besteht in der Erfüllung der Merkmale und Regeln, die vom formalen Typ (gemäss Einteilung in 1.1.1) her zu erwarten sind. Bezogen auf das Beispiel des relativ komplexen formalen Typs "syntaktisch organisiertes Wörterbuch" besteht die formale Adäquatheit in der formalen Konsistenz seiner Syntax als geregeltes System von zugelassenen Kategorienabfolgen und deren Verknüpfung. Dazu müssen folgende Bedingungen erfüllt sein:

1. Die Wörterbucheinträge müssen sich in elementare Teile zerlegen lassen, aus denen sich auch andere syntaktisch korrekte Einträge bilden lassen.
2. Diese elementaren Symbole müssen sich in einer endlichen Menge klassifizieren lassen.

3. Die elementaren Symbole müssen in der Wörterbuchsyntax als terminale Symbole auftreten können.
4. Die Klassen der terminalen Symbole repräsentieren in der Wörterbuchsyntax die präterminalen Kategorien des Regelsystems.
5. Im Regelsystem ist eine endliche Menge von Konstituentengruppen definiert.
6. Ein terminales Symbol, eine Konstituentengruppe oder mehrere Gruppen können ein Eintragungsschema korrekt beschreiben.

Diese formalen Syntax-Bedingungen beziehen sich auf alle linguistischen Ebenen eines Eintragungsschemas, insbesondere auch auf die semantische Ebene. Um z.B. den in Figur 1 gezeigten Ausschnitt eines semantischen Lexikoneintrags im oben definierten Sinne als korrekt zu akzeptieren, müssten die weiter unten genannten Syntax-Bedingungen erfüllt sein.

```

herstellen
WORTART: Verb
BEDEUTUNG 1: produzieren
  OBLIGATORISCH 1
    KASUSFUNKTION: Nomen (Nominativ)
    TIEFENKASUS   : Agent
    --SELEKTION   : Hum
  OBLIGATORISCH 2
    KASUSFUNKTION: Nomen (Akkusativ)
    TIEFENKASUS   : Object
    --SELEKTION   : Ina
  FAKULTATIV 1
    KASUSFUNKTION: Präposition (aus)
                  Nomen (Dativ)
    TIEFENKASUS   : Source
    --SELEKTION   : Mat
    TIEFENKASUS   : Location
    --SELEKTION   : Loc
BEDEUTUNG 2: schaffen
            etc.

```

Figure 3. 'Ausschnitt Lexikoneintrag (EVAR)': Müller81, p.65

Die Bedingungen:

1. Der Eintrag lässt sich in elementare Teile zerlegen:

- a) herstellen : LEMMA
- b) Verb : WORDART
- c) produzieren: Paraphrase der BEDEUTUNGSVARIANTE
- d) Nomen : AKTANT (Valenztheorie nach (Helbig69))
- e) (Nominativ): Syntaktische REALISIERUNG des Aktanten
- f) Agent : TIEFENKASUS (nach (Fillmore71))
- g) Hum : Semantisches SELEKTIONSMERKMAL des TK

h) Nomen : AKTANT
 i) (Akkusativ): Syn. REALISIERUNG
 k) Object : TIEFENKASUS
 l) Ina : SEMANTISCHES MERKMAL des Tiefenkasus
 etc.

Aus solchen und ähnlichen Elementen liessen sich weitere korrekte Einträge aufbauen.

2. Diese Elemente sind klassifizierbar: Es gibt endliche Mengen von Lemmata, Wortarten, Bedeutungsvarianten, Aktanten, syntaktischen Aktantenrealisierungen, Tiefenkasus, semantischen Merkmalen.

3. Die in 1. genannten Elemente sind terminale Symbole der Wörterbuchsyntax.

4. Die in 2. genannten Klassen sind präterminale Symbole der Konstituenten der Wörterbuchsyntax:

LEMMA ist präterminales Symbol der Konstituente EINTRAG.
 WORTART ist präterminal bezüglich MORPHOSYNTAKTISCHE Konstituente.
 BEDEUTUNGSVARIANTE bezüglich SEMANTISCHE Konstituente.
 AKTANT bezüglich VALENZRAHMEN.
 SYNTAKTISCHE AKTANTENREALISIERUNG ebenfalls bezüglich VALENZRAHMEN.
 TIEFENKASUS bezüglich KASUSRAHMEN.
 SEMANTISCHES MERKMAL bezüglich SEMANTISCHE Konstituente.

etc.

5. Ob es im Regelsystem der Wörterbuchsyntax des Lexikonsystems von EVAR eine endliche Anzahl von Konstituenten gibt, ist aus diesem Beispiel nicht ableitbar. (Die Syntax ist beschrieben in (Ehrlich86)).

6. Der oben dargestellte Eintrags "herstellen" bildet eine syntaktisch korrekte Konstituierung.

*

Wissenstheoretische Adäquatheit.

Die formale Repräsentation der Eintragsstruktur eines maschinenlesbaren Wörterbuchs als syntaktische und textgrammatische Struktur wird ergänzt durch den Formalismus zur Wissensrepräsentation des im Wörterbuch enthaltenen lexikalischen Wissens. Die Charakterisierung des Wörterbuchs als Wissensbasis komplementiert den syntaktischen Aspekt des Lexikons. Das Lexikon ist nicht nur ein syntaktisch geregeltes Gebilde, sondern hat als Ganzes auch eine Semantik, nämlich die Repräsentation von lexikalischem Wissen.

Dieser wissenstheoretische Aspekt des Lexikons ist weniger in der Wörterbuchsyntax realisiert als vielmehr im Formalismus zur Wissensrepräsentation. Diese Repräsentation ist nicht so offensichtlich, da das Wissen aus einer lexikalischen Wissensbasis für die Benutzung in einer Komponente des zugeordneten Sprachdaten verarbeitenden Systems zuerst repräsentationell aufbereitet wird. Ein Beispiel dafür sind Kasusinformationen. Im Lexikon ist nur die Kasuspotenz verzeichnet, nicht die aktuelle repräsentationelle

Lesart. Das trifft nicht nur auf Kasusinformationen zu. Lexikonwissen ist im allgemeinen vom aktuellen Kontext gelöst und deshalb nur potentielles, d.h. aktualisierbares Wissen. (Solches Wissen kann als *dynamisch bezeichnet werden). Das trifft auch bei morphologischen Angaben zu. Eine Flexionsklassenangabe beispielsweise zu einem Substantivlemma weist nur auf ein Set von Wortformen hin, d.h. die Angabe ist vom wissenstheoretischen Aspekt her gesehen noch nicht spezifisch. Ein weiteres Beispiel wären Fugenmorpheme. Werden einem Substantivlemma mehrere Fugenmorpheme zur Kompositabildung zugeordnet, bedeutet dies "lediglich" die Fügungspotenz des betreffenden Lemmas.

Für Vollformenlexika muss das eben Gesagte etwas eingeschränkt werden. Man könnte meinen, dass Vollformen spezifischere Wissensträger sind als Grundformen. Aber auch bei Vollformenlexika ist nicht das aktualisierte lexikalische Wissen repräsentiert, sondern nur die verschiedenen Möglichkeiten der Aktualisierung. Beispielsweise können auch Vollformen mehrere Bedeutungen haben. In morphologischer Hinsicht sind Vollformen dagegen oft spezifischer.

Um nun den repräsentationellen Aspekt des Lexikons zu beschreiben, benötigt man einen Repräsentationsformalismus, welcher der oben erwähnten speziellen Art von lexikalischem Wissen gerecht wird. Im Unterschied zu Knowledge Representation von spezifischem Wissen, wie z.B. Wissen über Fakten oder Wissen über eine Story oder Wissen über technische Modelle, muss der Formalismus zur Repräsentation von lexikalischem Wissen der Tatsache Rechnung tragen, dass es sich beim Lexikonwissen hauptsächlich um potentielles Wissen handelt, welches erst noch spezifiziert werden soll.

Die Spezifizierung kann auf unterschiedliche Art und Weise erfolgen. Eine Art der Spezifizierung ist die Kontextspezifizierung. Das ist aber nicht die einzige Art. Abgesehen vom Kontext der Wortform spielt auch der weitere Kontext eine Rolle, nicht zuletzt auch der pragmatische. Hinzu kommen noch die subtileren Aktualisierungen, die z.B. auf Ironie oder anderen inszenierten Aktualisierungsverschiebungen beruhen. Lexikalische Repräsentationsformalismen müssen also speziell geeignet sein, potentielles Wissen so zu repräsentieren, dass die Repräsentation bei der Aktualisierung auch verwendet werden kann.

Derartige Repräsentationsformalismen gibt es sehr viele, nur werden sie nicht immer so genannt. Gemeinsam ist diesen Formalismen, dass das, was sie repräsentieren, erst nach einer geeigneten Aufbereitung vereindeutigt wird. Vielleicht ist es übertrieben zu behaupten, lexikalische Einträge seien nicht als eindeutige Repräsentationen zu verstehen. Aber paradox könnte man formulieren, dass gerade durch eine möglichst vieldeutige Repräsentation die Möglichkeit zur Eindeutigkeit geschaffen wird.

Die Beschreibung von lexikalischem Wissen mit einem speziellen Repräsentationsformalismus erfolgte bereits durch den Pionier auf dem Gebiet der semantischen Netzwerke:

"The idea of a semantic network representation for human knowledge is generally acknowledged to have originated in the work of Quillian (1966, 1967, 1968, 1969; Bell and Quillian, 1977); Quillian proposed an associational network model of 'semantic memory' in his Ph.D. theses in 1966. His intent was to capture in a formal representation the 'objective' part of the meanings of words

so that 'humanlike use of those meanings' would be possible (1966,p.1). The representation was composed of nodes, interconnected by various kinds of associative links, and closely reflected the organization of an ordinary dictionary. The nodes were to be considered 'word concepts', and links from a concept node pointed to other word concepts, which together made up a definition, just as dictionary definitions are constructed from sequences of words defined elsewhere in the same volume. The structure thus ultimately became an interwoven network of nodes and links."

(Brachman79p.5/6)

Besonders relevant in Bezug auf den speziellen Charakter des lexikalischen Wissens ist die Formulierung

"to capture the 'objective' part of the meanings of words so that 'humanlike use of those meanings' would be possible"

Das Lexikon oder eben das semantische Netz ist das objektivierbare Bedeutungspotential, mit dem der aktuelle Gebrauch der Wörter in Sätzen oder Texten semantisch spezifizierbar wird. Objektivierbar wird dieses Potential deshalb genannt, weil im semantischen Netz alle möglichen Bedeutungen objektiv repräsentiert sein sollen. Im jeweils subjektiven Gebrauch der Wörter ("humanlike") sind die potentiellen Bedeutungsrepräsentation auf wenige aktuelle Varianten restringiert.

Es sollte aber klar sein, dass Quillians an einem definitonischen Bedeutungswörterbuch orientierter Netzformalismus nur ein Beispiel für viele weitere lexikonspezifische Repräsentationsformalismen ist. Um auf der von Quillian gewählten semantischen Ebene zu bleiben, könnte man an weiteren lexikonspezifischen semantischen Repräsentationsformalismen folgende nennen: Thesaurus, Valenz, Tiefenkasus und logische Prädikate. Gemeinsam ist diesen Formalismen und dem Netzwerkformalismus von Quillian, dass sie dazu dienen, eine möglichst grosse Zahl von Bedeutungspotenzen der eingetragenen Lemmata so differenziert abzubilden, dass mit Hilfe dieser Differenzierungen ein aktuell in einem Text realisiertes Lexem spezifiziert werden kann. (Unter Zuhilfenahme von weiteren Wissensquellen, wie etwa der Resultate einer vorangehenden Syntaxanalyse oder bereits erschlossener Diskursinhalte.)

Es wäre aber falsch anzunehmen, dass Repräsentationen von potentielllem Wissen nur auf der semantischen Ebene der Lexikoneinträge zu finden sind. So sind z.B. auch Wortklassenangaben als solche Repräsentationen zu betrachten, denn auch sie bilden ein Potential, das bei der Aktualisierung der Wortform spezifizierend wirkt.

Infolge der Beschreibung der lexikalischen Repräsentation als besonderer Form der Wissensrepräsentation können viele der in der KI erarbeiteten Merkmale solcher Formalismen für die Belange der lexikalischen Repräsentation adaptiert werden. Insbesondere lassen sich Überlegungen zur Differenzierung des Repräsentationsformalismus auf verschiedene Darstellungsebenen praktisch vollständig übernehmen, wobei lediglich daran erinnert werden muss, dass lexikalisches Wissen nicht Fakten, sondern Potentiale repräsentiert. Die formale Repräsentation von Wissen kann nach (Brachman79,p.31) in folgenden von ihm in dieser Reihenfolge genannten Darstellungsebenen realisiert werden:

implementational level
logical level
epistemological level
conceptual level,
linguistic level

Er bezieht sich bei diesen Unterscheidungen zwar auf semantische Netzwerke, aber sie lassen sich auch auf Wörterbücher anwenden. Diese Unterscheidungen traf Brachman in kritischer Auseinandersetzung mit diversen früher vorgeschlagenen Repräsentationsformalismen, u.a. den semantischen Netzwerken Quillians. Brachman stellte dabei fest, dass die aufgezählten Repräsentationsebenen z.T. vermengt und ignoriert wurden.

Was können nun diese Ebenen im Hinblick auf Wörterbücher bedeuten?

"Linguistic level" exemplifiziert Brachman an dieser Stelle mit "arbitrary concepts, words, expressions". "Linguistic level" ist also hier im Sinne von linguistischer Oberflächenstruktur verstanden und hat nur insofern mit der repräsentationellen Strukturierung des lexikalischen Sprachwissens zu tun, als repräsentationelle Darstellungen selber auch sprachliche Darstellungen sind.

"Conceptual level" exemplifiziert Brachmann mit

"semantic or conceptual relations (cases), primitive objects and actions".

Kategorien wie Kasusrahmenkonzepte mit ihren Argumenten und semantische Primitiva im nominalen und verbalen Bereich sind bereits konstitutive Teile des Repräsentationsformalismus für (lexikalische) Wissensbasen. Allerdings stellen sie noch nicht diesen Formalismus selbst dar, sondern sind nur Teile davon.

Die Gesamtheit dieser eben erwähnten Konstrukte des "conceptual level" sind auf der epistemologischen (wissenstheoretischen, erkenntnistheoretischen) Ebene organisiert und bilden erst in ihrer organisierten Gesamtheit einen Repräsentationsformalismus. Bezogen auf Wörterbücher drückt die Unterscheidung von conceptual und epistemological level in etwa den Unterschied zwischen lexikalischen Kategorien (wie Wortklasse, Valenz, Tiefenkasus) und deren Organisation in einer Gesamtform lexikalischer Wissensrepräsentation aus. Brachmans spezielle Aufmerksamkeit gilt gerade der epistemologischen Ebene: Die Organisation konzeptioneller Konstrukte in einen einheitlichen, grösseren Formalismus mit limitierten und definierten Strukturierungsmitteln bildet den Kern eines Formalismus für Wissensrepräsentation. Auf dieser Ebene kann eine Repräsentationssprache definiert werden, was Brachman mit KL-ONE auch getan hat (Brachman85, p171ff).

Das Ansetzen eines "logical level" basiert auf der stillschweigenden Voraussetzung, dass

"sich jedes Wissen geeignet nach den Prinzipien der Prädikatenlogik sinnvoll faktorisieren (also in formale Einheiten zerlegen) lässt". (Brietzmann84a, p.39).

In Bezug auf lexikalisches Wissen bedeutet dies, dass eine Logik der Potentialität von Lexemrepräsentationen in Bezug auf Prädikate konstruiert werden muss: Das Lexikon als Wissensquelle für die Konstruktion logischer Prädikate im Text.

Auf die Implementationsebene gehe ich in diesem Zusammenhang nicht näher ein. Eine Implementation des von Brachman und anderen entwickelten Repräsentationsformalismus findet sich aber im Semantik- und Pragmatikmodul des Spracherkennungssystems EVAR. Brachman weist darauf hin, dass die Implementationsebene nicht mit den anderen Ebenen vermischt werden soll.

Obwohl Brachman diese 5 Betrachtungsebenen auf strukturierte semantische Netzwerke angewandt hat, scheint es sinnvoll, diese Unterscheidungen auch auf maschinenlesbare Wörterbücher anzuwenden. Diese Betrachtungsweise bestätigt sich, wenn man die ursprüngliche Formulierung der Idee semantischer Netzwerke durch Quillian im Auge hat, auf die Brachman selbst hinweist.

Nach dieser knappen Vorstellung der wissenstheoretischen Implikationen von Repräsentationsformalismen für Wörterbuchwissen gilt es, einer Verwechslung mit einer anderen Ebenen-Unterscheidung vorzubeugen, welche nicht mit dem Repräsentationsformalismus, sondern mit dem zu Repräsentierenden zusammenhängt.

Linguistische Adäquatheit.

Nach linguistischen Gesichtspunkten lässt sich lexikalisches Wissen ebenfalls in Stufen einteilen. Eine vereinfachende Sichtweise besteht darin, lexikalisches Wissen von regelhaftem Wissen zu trennen. Diese herkömmliche Opposition basiert auf dem in der TG aufgebauten Gegensatz zwischen Lexikon und Grammatik. Eine solche Sichtweise verdeckt aber u.U. die Tatsachen, dass mit Regelwissen nicht nur die grammatikalische, syntaktische Ebene beschrieben werden kann, und dass lexikalisches Wissen selbst linguistisch vielschichtig ist. Das heisst aber nicht, dass die Opposition Lexikon - Regelwissen falsch ist. Vielmehr sollte diese Unterscheidung dahingehend ergänzt werden, dass sowohl Lexikon als auch Regelapparat je in sich geschichtet sind. Die Schichtungen in Lexikon und Regelapparat sind parallel. Diese parallelen Schichten sind:

Phonologische lexikalische Information und phonologisches Regelwissen.

Morphologische lexikalische Information und morphologisches Regelwissen.

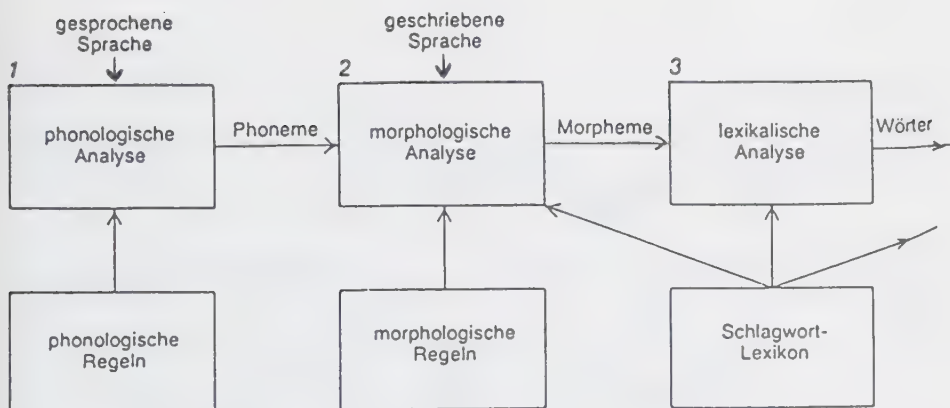
Syntaktische lexikalische Information und syntaktisches Regelwissen.

Semantische lexikalische Information und semantisches Regelwissen.

Logische lexikalische Information und logisches Regelwissen.

Pragmatische lexikalische Information und pragmatisches Regelwissen.

Je nach Differenzierungsgrad kann man mehr oder weniger von diesen Schichten unterscheiden. Als linguistische Standarderteilung kann eine von T. Winograd vorgeschlagene Stufung angesehen werden. Terry Winograd skizziert in (Winograd84, p.96-100) die folgenden sprachlichen Analysestufen mit den jeweils zugehörigen Wissensquellen:



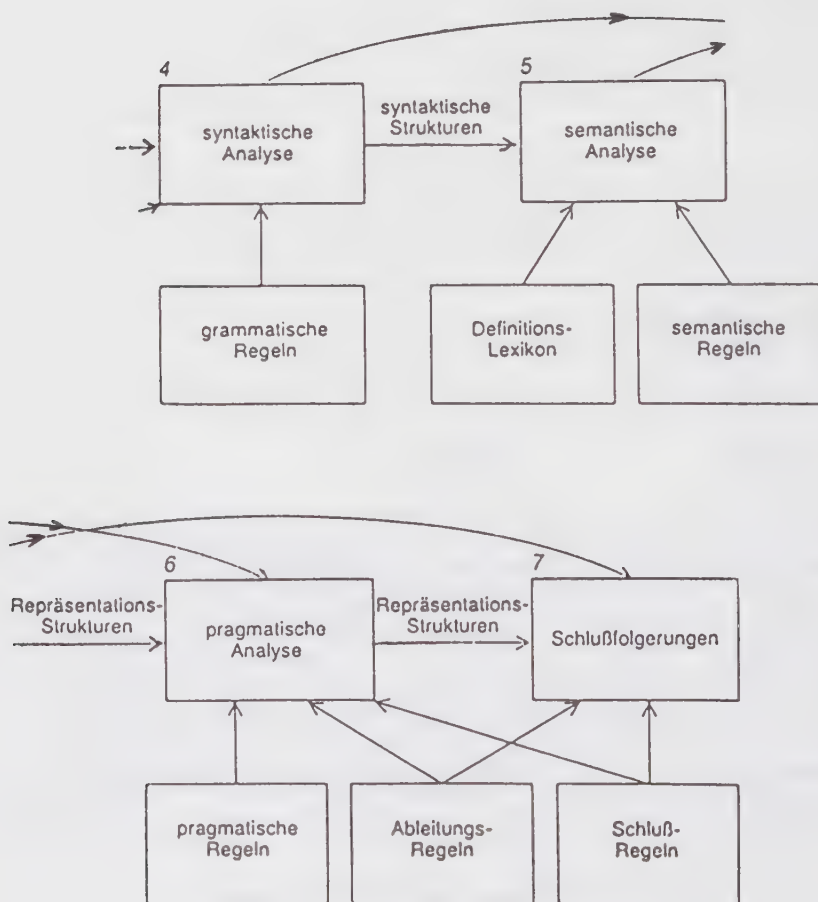


Figure 4. 'Analyseprozesse und Wissensquellen': nach Winograd

Die numerierte obere Reihe der Boxen entspricht den Analyseprozessen, die untere Reihe der mit Pfeilen der oberen zugeordneten Kästen entspricht den Wissensquellen.

"Wenn die Sprache in gesprochener Form vorliegt, ist die erste Analyse phonologisch (1): Der Computer analysiert die Schallwellen.

Liegt die Sprache in geschriebener Form vor, ist die erste Analyse morphologisch (2): Der Computer zerlegt jedes Wort in seine Wurzel und die Flexionselemente (zum Beispiel "geh-st").

Dann folgt die lexikalische Analyse (3), in der der Computer die Wörter ihrer lexikalischen Kategorie zuweist (zum Beispiel Substantiv) und 'Merkmale' wie etwa Plural identifiziert.

Es folgt die syntaktische Analyse, das sogenannte Parsen (4): die Anwendung der grammatischen Regeln zum Erschliessen der Satzstruktur.

Danach kommt die semantische Analyse (5). Hier wird der Satz in eine Form überführt, die als Basis für Schlussfolgerungen geeignet ist.

Die letzte Stufe ist pragmatisch (6): Sie macht den Kontext des Satzes deutlich, so die Beziehung zwischen der Zeit, zu der er geäußert wurde, und der Zeit, auf die er sich bezieht. Der Computer kann daraufhin Schlussfolgerungen ziehen (7), vielleicht um die Beantwortung eines Satzes vorzubereiten. Die heutigen Programme basieren auf vereinfachten Sprach- und Bedeutungssystemen." (Winograd84, p. 96/97)

Wie man sieht, geht es hier nicht wie bei Brachmann, um die Abgrenzung der Repräsentationsebenen voneinander, z.B. der epistemologischen Repräsentationsebene von der konzeptionellen und der logischen Ebene, sondern um die inhaltlich differenzierte Zuordnung von nach linguistischen Kriterien geschiedenen Wissensquellen zu Analyseprozessen. Diesen differenzierten Zuordnungen ist eine linguistische Level-Unterscheidung zugrundegelegt, die nicht die Logik des Repräsentationsformalismus betrifft, sondern die Logik der Linguistik.

Gemeinsam ist beiden Betrachtungsweisen aber das Verständnis von Wörterbüchern als Wissensbasen. Sehr klar werden auch bei Winograd Wörterbücher als Wissensbasen dargestellt. Winograd macht dabei auch auf die Parallelität von unmittelbar einem Prozess zugeordneten Regelapparaten und Wissensbasen aufmerksam. Solche, einem Prozess zugeordnete Regelapparate haben nichts mit den die Wissensbasen selber strukturierenden Regelapparaten zu tun, sondern sind selber als Wissensbasen für den zugeordneten Prozess zu betrachten. So betrachtete regelbasierte Wissensbasen sind zusammen mit lexikalischen Wissensbasen als gemeinsame Quellen zu zugeordneten Prozessen zu betrachten. Beispielsweise bestehen diese Gemeinsamkeiten zwischen einem Definitionslexikon und semantischen Regeln: Beides sind Wissensbasen in Bezug auf den Analyseprozess "semantische Analyse".

*

Zusammenfassend kann gesagt werden, dass bei der Abschätzung der Adäquatheit lexikographischer Abbildungen in Wörterbüchern

- die formalen (die Lexikonsyntaktik betreffenden),
- die wissenstheoretischen (den Repräsentationsformalismus betreffenden) und
- die linguistischen (die Sprachebenen betreffenden) Aspekte berücksichtigt werden sollten.

1.2 * LEMMA UND EINTRAG.

In Abschnitt 1.1 wurde die Differenzierung des Begriffs "Wort" bei der Herausarbeitung der formalen Komplexitätsstufen der Eintragsstruktur eines Wörterbuchs und der Formulierung von darauf bezogenen Adäquatheitsbedingungen nicht benötigt, weil es zur formalen Unterscheidung der Komplexitätsstufen genügte, ein Wort vereinfachend als Element eines Abfolgeschemas zu betrachten. Der Untersuchungsgegenstand war die gesamte Eintragsstruktur. Im nun folgenden Abschnitt geht es um den Eintrag selbst oder, bildlich gesprochen, um eine Verengung der Perspektive auf die "Binnenstruktur" des Eintrags. Zur Durchleuchtung der Binnenstruktur eines Eintrags ist aber gerade die Differenzierung des Begriffs "Wort" erforderlich.

In der Lexikographie wird die Einheit, welche den Eintrag bzw. den Wörterbuchartikel repräsentiert, Suchwort, Stichwort oder Lemma genannt. Lemma ist der geeignetere Ausdruck, weil das Stich-"wort" gar kein Wort zu sein braucht.

"Der obligatorische Bauteil Lemma (vgl. Wiegand83) kann in ein- und demselben Wörterbuch einen sehr unterschiedlichen lexikologischen Status haben. Lemmafähig ist jede lexikalische Einheit. Es kommen in Frage: Wörter (auch Ableitungen, Zusammensetzungen), Wortformen (z.B. 'ging', 'bist'), Affixe (Suffixe, Präfixe, Infixe), Affixoide (z.B. -muffel in 'Gurtenmuffel'), weitere Wortbildungselemente, Wurzeln (z.B. frz. -tim- wegen 'timoré', 'timide', 'intimider'), Wortgruppenlexeme (z.B. d'émblée), Teile von Wortgruppenlexemen ohne Monemstatus (z.B. 'Bockshorn' aus 'jemanden ins Bockshorn jagen'; 'émblée' aus 'd'émblée'), Abkürzungen (z.B. Z = Zeile), Eigennamen." (Hausmann85, p.372)

Dieser Tatsache trägt z.B. das Bedeutungswörterbuch Duden (Bd 10, 1985) dadurch Rechnung, dass dem eigentlichen Wörterbuch eine

"Liste der Wortbildungsmittel, die innerhalb der alphabetischen Abfolge als Stichwörter [! H.H] erscheinen"

vorgeordnet ist (p. 16-18). Motiviert wird die Aufnahme von über 400 Wortbildungsartikeln mit der Produktivität der Sprache. Dabei werden insbesondere auch Affixoide (= Halbpräfixe bzw. Halbsuffixe) berücksichtigt, und die Dudenredaktion erwähnt den Unterschied zur Komposita-Bildung, die ja für die Wortbildung im Deutschen eine sehr grosse Rolle spielt.

Abstrakter formuliert und erweitert, kann nun "Lemma" so definiert werden:

"Ein Lemma ist der Repräsentant einer Menge morphologisch und semantisch verwandter Types. Es wird zumeist in der für seine Wortklasse üblichen Grundform notiert."

"Eine Vollform wird lemmatisiert, indem sie deflektiert und -falls mehrdeutig- disambiguiert wird."
(Drewek82)

Diese Definition ist dynamisch, weil sie die Richtung weist, wie lemmafähige Einheiten zu ermitteln sind:

1. Ermittlung von Types.
2. Morphologische und semantische "Verwandtschaft" erkennen.
3. Disambiguierung.
4. Deflexion.
5. Notierung der Grundform.

Es geht also darum, die Generierung einer eindeutigen (disambiguierten) und morphologisch normierten (deflektierten) Grundform als obligatorischer Basis (Lemma) eines Wörterbucheintrages zu präzisieren.

1.2.1 Eindeutigkeit.

In der linguistischen Datenverarbeitung wurden Algorithmen entwickelt, welche zur Lemmatisierung eingesetzt werden können. Zu nennen sind hier u.a. die Verfahren von R. Dietrich (Dietrich73), das Programm LEMMA und dessen Weiterentwicklung (Willée79, Willée84), das Programm SALEM aus dem Kreis des Sonderforschungsbereichs 100 in Saarbrücken (Eggers80) und das Programm LEM von R. Drewek (Drewek80/83). [**16]

Lemmatisierungsprogramme können grundsätzlich in zwei Richtungen angewendet werden. Zum einen können mit Hilfe eines beigeordneten Wörterbuchs in den zu analysierenden Texten Lemmata identifiziert werden, und zum andern kann mittels umfangreicher Textkorpora der Grundstock von Wörterbüchern -das Lemmainventar- hergestellt werden.

Die "einfachere" und vorläufig zuverlässigere Art von Lemmatisierungsprogrammen ist interaktiv konzipiert, d.h. es sind nicht alle Schritte der Lemmatisierung algorithmisiert. Für das Programm LEM von R. Drewek gilt folgende Arbeitsorganisation:

1. Mittels des Programms KWI (Drewek75ff) wird ein Index der Wortformen über dem Text erstellt. (Die Texttokenmenge wird in eine sortierte Folge der Types mit Fundstellenreferenz transformiert).

2. Der Index wird mit den Lemmata eines Wörterbuchs abgeglichen. Bei erfüllter Matchbedingung werden aus dem Wörterbuch einer oder mehrere Lemmavorschläge generiert (und auf dem Schirm angezeigt).

3. Der Kontext des aktuellen Tokens (in Bezug auf das zu bearbeitende Type) ist ebenfalls auf dem Schirm sichtbar, weil im Index die Textbelegstellen markiert sind. (Konkordanzzeilen identischer aktueller Wortformen).

4. Der Bearbeiter entscheidet aufgrund des Kontextes, des pragmatischen Ziels seiner Lemmatisierung und der die Matchbedingung erfüllenden Lemmavorschläge aus dem Wörterbuch, welcher Lemmakandidat zum Token im Kontext passt. Falls das Wörterbuch bereits mit vielen Lemmata ausgestattet ist, ist dieses Verfahren effizient. Falls das Wörterbuch leer ist, wird die oben erwähnte umgekehrte Bearbeitungsrichtung eingeschlagen und das Wörterbuch erstellt. Dieser zweite Weg ist aufwendig, weil vom Bearbeiter die Grundform ermittelt werden muss, und die morphologischen, semantischen und sonst noch gewünschten Merkmale zu dieser Grundform erfasst werden müssen.

5. Als Output des Verfahrens kann aufgrund der nun lemmatisierten Texttypes eine lemmatisierte Konkordanz hergestellt werden, welche die Textbelegstellen nicht nach Wortformen sortiert, sondern nach Lemmata sortiert aufweist.

Lemmatisierung ist ein sehr komplexer Prozess und wird mit folgender Formel vereinfacht dargestellt:

LEMMA = DEFLEXION (DISAMBIGUIERUNG (TYPE))
(Drewek80/83)

Zuerst erfolgt die Funktion TYPE, dann DISAMBIGUIERUNG, dann DEFLEXION. Diese Reihenfolge ist wesentlich, weil Types mit identischer Wortform aber unterschiedlicher Bedeutung in manchen Fällen verschieden deflektiert werden.

Beispiel: (das graphematische Unterscheidungskriterium der Grossschreibung wird hier ignoriert)

FLOH wird disambiguiert zu 1. Floh, subst. und
2. fliehen, verb,

Nebenbei sei hier angemerkt, dass die beiden erzeugten Lemmata als Wortformen in der grammatischen Form immer noch mehrdeutig sind:

- Floh (subst) repräsentiert mehrere Kasus.
- floh, (finite Verbform) repräsentiert mehrere Personen.
- fliehen (infinite Verbform) fällt mit finiten Formen zusammen.

Diese Ambiguitäten komplizieren die Lemmatisierung, weil immer darauf geachtet werden muss, dass man die Disambiguierung auf deflektierter Ebene nicht mit grammatikalischer Ambiguität verwechselt.

Die Funktion DISAMBIGUIERUNG impliziert Wissensbasierung aus verschiedenen Wissensquellen. Im oben beschriebenen interaktiven Lemmatisierungsprozess können diese Quellen wie folgt lokalisiert werden:

1. Die Kotextmenge des Types (d.h. die aktuellen Konkordanzzeilen, wobei nicht festgelegt ist, wie breit diese Zeilen sind).
2. Die Lemmavorschläge (inklusive codierte Einträge).
3. Die Kriterien des Matchingalgorithmus zwischen aktuellem Type und Lemma.
4. Die Auswahlkriterien des Bearbeiters.
5. Das umfangreiche Sprachwissen des Bearbeiters, falls kein Vorschlag passt.

1.2.2 Disambiguierung.

Versucht man, den ganzen Prozess zu automatisieren, muss berücksichtigt werden:

1. Der Matchingalgorithmus soll alle möglichen Lemmata liefern.
2. Die gelieferten Lemmavorschläge müssen soviel Information aufweisen, wie für die Disambiguierung benötigt wird.
3. Der Auswahlprozess muss über den Kotexten operieren können.

Um die Tragweite dieser drei Anforderungen abzuschätzen und die Schwierigkeiten zu minimieren, können unterschiedliche Arten von Ambiguitäten isoliert werden, die dann in unterschiedlichen Disambiguierungsfunktionen behandelt werden können. T. Winograd (Winograd84, p.90) unterscheidet 5 verschiedene Arten von Ambiguität:

"1. Bei der einfachsten Art der Ambiguität, bekannt als lexikalische Ambiguität, besitzt ein Wort mehr als nur eine mögliche Bedeutung. Beispielsweise kann der Satz 'das Schloss wurde im 16. Jahrhundert gebaut' sich sowohl auf ein Gebäude als auch auf eine Schliessvorrichtung beziehen."

"2. Bei der strukturellen Ambiguität geht das Problem über das einzelne Wort hinaus. Als Beispiel dient der Satz 'Ede sah den Mann mit dem Fernglas'. Der Satz hat zwei Interpretationen, je nachdem ob sich 'mit Fernglas' auf 'den Mann' oder auf 'sah' bezieht. Daher besitzt der Satz zwei mögliche grammatische Strukturen unter denen man wählen muss."

"3. Noch subtiler ist eine Ambiguität der Tiefenstruktur: Zwei Lesarten können an der Oberfläche die gleiche grammatische Struktur besitzen, aber dennoch sich in ihrer Bedeutung unterscheiden. (...) 'The chickens are ready to eat' ; 'Die Hühner sind bereit zu fressen' oder 'Die Hühner sind zum Essen fertig'. Ambiguität der Tiefenstruktur zeigt auch der deutsche Satz 'Ede ist fertig zum Photographieren'. Er impliziert, dass jemand jemanden photographieren soll, aber welche Rolle spielt dabei Ede?"

"4. Eine vierte Art der Ambiguität -die semantische- (...) ergibt sich, wenn ein Ausdruck unterschiedliche Rollen in der

Gesamtbedeutung eines Satzes spielen kann. Der Satz 'Ede will eine Norwegerin heiraten' ist ein Beispiel. In einer Lesart wird der Ausdruck 'eine Norwegerin' referentiell gebraucht: Ede möchte eine bestimmte Person heiraten, und der Sprecher des Satzes hat eine Eigenschaft dieser Person -dass sie aus Norwegen kommt- gewählt, um sie zu bezeichnen. In einer anderen Lesart dieses Satzes ist der Ausdruck attributiv. Weder Ede noch der Sprecher haben eine bestimmte Person im Auge; der Satz bedeutet einfach, dass Ede sich wünscht, jemanden mit norwegischer Nationalität zu ehelichen."

"5. Eine fünfte Art der Ambiguität kann pragmatisch genannt werden. Sie tritt bei Verwendung von Pronomina und anderen verkürzten Ausdrücken wie 'einer' auf. Ein Beispiel ist der Satz 'Wenn ein heller Mond einen dunklen Tag beschliesst, wird ein hellerer folgen'. Ein hellerer Tag oder ein hellerer Mond?"

Diese Unterscheidung von lexikalischer, struktureller, tiefenstruktureller, referenzsemantischer und pragmatischer Ambiguität impliziert für jede Art eine eigene Disambiguierungsfunktion. Und jede Disambiguierungsfunktion impliziert eine zugeordnete Wissensbasis. In diesem Kapitel (1.2) geht es aber nicht um den Entwurf von Disambiguierungsfunktionen, sondern um die Fundierung des Lexikoneintrags im Wissen um eben diese Funktionen.

1.2.3 Informationslinguistisch fundierter Eintrag.

Ein Lexikoneintrag eines maschinenlesbaren Wörterbuchs sollte formal, epistemologisch und lexikographisch fundiert sein. Formal fundiert ist ein Eintrag, wenn er in die Lexikonsyntax passt, epistemologisch fundiert ist er, wenn er nach den Vereinbarungen des verwendeten Repräsentationsformalismus Wissen darstellt und lexikographisch fundiert ist er, wenn dieses Wissen linguistisch korrekt ist.

Was bedeutet in diesem Zusammenhang linguistisch korrekt?

Diese Frage soll hier nur im Zusammenhang mit den semantischen Kategorien eines Eintrag diskutiert werden.

Die semantischen Einträge sollen die semantische Potenz des betreffenden Lemmas repräsentieren. Darunter ist zu verstehen, dass die Lexikoneinträge möglichst viele semantische Aktualisierungsmöglichkeiten enthalten sollen, die in den Funktionen der semantischen Komponente des Systems berücksichtigt werden. Solche Funktionen sind:

- a) Semantische Verträglichkeiten überprüfen.
- b) Semantische Erwartungsrahmen aufbauen.
- c) Semantische Rollen in einem Satz identifizieren.
- d) Syntaktische Strukturen semantisch interpretieren.

e) Semantische Strukturen für die logische und pragmatische Interpretation aufbereiten.

f) Ambiguitäten auflösen, wobei hier im Sinne Winograds differenziert werden kann:

- Die Auflösung lexikalischer Ambiguität (Auswahl der für den laufenden Satz passenden Bedeutungsvariante der Wortform.

- Die Auflösung struktureller Ambiguität (sie ist identisch mit der oben genannten Funktion der semantischen Interpretation syntaktischer Strukturen).

- Die Auflösung tiefenstruktureller Ambiguität (sie basiert auf der ebenfalls schon erwähnten Identifikation semantischer Rollen im Satz).

- Die Auflösung referenzsemantischer Ambiguität (erfordert mindestens die Berücksichtigung des Kontexts, wenn nicht des Kontexts).

- Das Gleiche gilt für die Auflösung pragmatischer Ambiguität.

Wie man sieht, ist die Funktion der Disambiguierung mit den anderen semantischen Funktionen verknüpft.

g) Eine weitere semantische Funktion ist es, bedeutungsgleiche Sätze oder Ausdrücke (Paraphrasen) zu erkennen.

Es könnten noch mehr semantische Funktionen genannt werden, aber zur Bestimmung der semantischen Potenz der Lexikoneinträge genügen diese.

Nehmen wir als Beispiel die Disambiguierung syntaktischer Strukturen:

Damit Disambiguierung stattfinden kann, muss bei der Analyse eine Auswahl möglicher Bedeutungsvarianten vorliegen. Diese Lesarten werden durch den syntaktischen Parser generiert. Die Disambiguierung, d.h. die Auswahl einer Lesart, beruht auf semantischen Verträglichkeitstests und der Abprüfung von Erwartungsrahmen mit Füllern.

Sowohl die semantischen Merkmale, auf denen die Verträglichkeitstests beruhen, als auch die Erwartungsrahmen werden aus dem Lexikoneintrag bezogen.

Noch ein Beispiel:

Die Disambiguierung mehrerer Lesarten einer Wortform. Auch diese Lesarten werden durch den Lexikonzugriff produziert. Im günstigsten Fall kann durch die syntaktische Analyse eine Lesart ausgewählt werden.

Diese beiden Beispiele zeigen, dass im Lexikoneintrag semantisches Lexikonwissen nicht nur zur Disambiguierung, sondern auch zur Produktion von Lesarten dient. Die semantische Potenz des Lexikoneintrags besteht gerade auch in der letztgenannten Eigenschaft.

Im informationslinguistisch fundierten Lexikoneintrag wird diese Dialektik der semantischen Potenz durch Einbau von Generatoren und von Reduktoren (Selektoren) von Bedeutungsvarianten realisiert. Der wichtigste Generator für Bedeutungsvarianten ist die möglichst zahlreiche und differenzierte Codierung von Homonymen, Polysemen und Varianten. Falls der Begriff 'Lemma' selbst schon auf Disambiguierung aufgebaut ist, wie dies in der in 1.2.1

eingeführten Lemmadefinition der Fall ist, würde diese Differenzierung in der Lemmageneration bestehen, allenfalls auch in der Variantengeneration zu einem Lemma. Dies gilt für alle Wortarten.

Als Reduktoren (Selektoren) kommen vor allem Erwartungsrahmenkonstrukte in Frage, wie Valenzrahmen und Kasusrahmen. Diese sind nicht nur auf Verben beschränkt, sondern finden sich auch bei Adjektiven, Substantiven, Funktionswörtern.

Zusammenfassend kann gesagt werden, dass der informationslinguistisch fundierte Eintrag potentiell Wissen enthält, welches sowohl der Produktion als auch der Reduktion von Lesarten dient.

Bei der Generierung informationslinguistisch fundierter Lexikoneinträge müssen abgesehen von den eben zusammengefassten Hauptmerkmalen solcher Einträge auch Differenzierungen berücksichtigt werden, die damit zusammenhängen, ob das Lexikon Vollformen und/oder Grundformen und/oder Stammformen enthält. Sind beispielsweise Vollformen und Grundformen gemischt vertreten, dann können in den Vollformen alle Informationen zu Eintragskategorien, die sich auf der Formenebene nicht ändern, durch Verweis auf die zugehörige Grundform behandelt werden. Bei diesem Beispiel handelt es sich um einen Spezialfall von sog. "Bündelung", welche in der lexikographischen Praxis auch bei maschinenlesbaren Wörterbüchern die Eintragsstruktur beeinflusst:

"Aus pragmatischen Gründen (Lesbarkeit, Speicherökonomie) werden deshalb in konventionellen und auch in maschinenlesbaren Lexika Klassen solcher Einträge, die sich nur in einer oder wenigen Komponenten unterscheiden, nach bestimmten Regeln zu einem Eintrag zusammengefasst. Mit dem Sinken der Zahl der Einträge steigt allerdings die erforderliche Informationsmenge innerhalb eines Eintrages ('Bündelung')."

(Hess83, p. 46)

Bei Hess et al. werden an dieser Stelle (p.46/47) fünf typische Beispiele genannt, wobei das bereits erwähnte Verfahren des Zusammenfassens mehrerer Vollformen Einträge ebenfalls angeführt wird:

"Vollform-Einträge bei gleichem Lemma (Grundform), gleicher Wortklasse und gleichen semantischen Angaben, aber mit unterschiedlichen Flexionsmorphemen und flexionsmorphologischen Angaben können in einem Eintrag zusammengefasst werden, wobei Wortlautbereich WG [Wortlaut wird im Eintragungsschema von Hess et al. als WG bezeichnet, vergl. Kapitel 2.2.5 H.H.] die Angabe des Lemmas enthält und im morphologischen Bereich Angaben zur gesamten Flexionsreihe folgen müssen.

Beispiele hierfür sind die Stammformeneinträge wie teilweise im SADAW". [zu SADAW vergl. Kapitel (2.2.4)].

(Hess83, p. 46)

1.3 * VERFLECHTUNGEN ZWISCHEN DER MIKRO- UND DER MAKROSTRUKTUR.

Nach der getrennten Behandlung von Makrostruktur (Kapitel 1.1) und Mikrostruktur (Kapitel 1.2) können jetzt die wichtigsten Verflechtungen dieser Strukturen analysiert werden. Um terminologische Klarheit über diese beiden Begriffe zu erzielen, eignet sich folgende Passage aus dem Handbuch der Lexikographie:

"Das Definitionswörterbuch des Typs 'Wörterbuch der deutschen Gegenwartssprache' besteht aus zwei Dimensionen:

a) der (alphabetisch) geordneten Folge von Wörterbucheinträgen (Lemmata), über die die im Wörterbuch vermittelte Information zugänglich ist. Diese Dimension des Wörterbuchs heisst Makrostruktur.

b) den über die Lemmata (als Leitwörtern) zugänglichen Wörterbuchartikeln als Trägern der im Wörterbuch vermittelten Information. Diese Dimension des Wörterbuchs heisst Mikrostruktur (wobei vorausgesetzt ist, dass alle Artikel eines Wörterbuchs gleich strukturiert sind)."

(Hausmann85, p.372)

Bei maschinenlesbaren Wörterbüchern trifft es allerdings nicht zu, dass die in den Artikeln gespeicherte Information nur über die alphabetisch geordnete Lemmafolge zugänglich ist. Das ändert aber nichts an der Tatsache, dass auch bei maschinenlesbaren Wörterbüchern die Mikro- und die Makrostruktur sinnvoll voneinander unterschieden werden können.

In Bezug auf diese beiden Begriffe nimmt Hausmann in einer Anmerkung durch eine Abgrenzung von (Wiegand83, p.431,453) eine präzisierende Erweiterung vor:

"Wiegand fasst diese Termini ausschliesslich im Sinne von Ordnungsstrukturen und nicht -wie wir im Interesse eines vereinfachten Sprechens über die Dinge- als Strukturen und Mengen. Die Mengen heissen bei Wiegand 'Lemmabestand' und 'Wörterbuchartikel' (bestehend aus lexikographischen Textbausteinen und Textelementen)."

Die Unterscheidung zwischen Mikro- und Makrostruktur eines Wörterbuchs basiert auf der Voraussetzung, den Wörterbuchartikel (auch Lemma, Eintrag genannt) isolieren zu können. Im allgemeinen ist das auch der Fall, und in gedruckten Wörterbüchern wird der Beginn des nächsten Artikels jeweils deutlich markiert.

Das sollte aber nicht darüber hinwegtäuschen, dass dies in vielen Fällen lediglich eine Fiktion ist. Gerade bei Einträgen in Bedeutungswörterbüchern wird das deutlich:

Frucht-bar-keits-zau-ber <m.; bei Naturvölkern>
kult. Bräuche zur Erhaltung u. Stärkung der
Fruchtbarkeit (des Bodens, Viehs, auch der Menschen)

In diesem Beispiel aus (Wahrig85) sind fast sämtliche Wörter in der Paraphrase lediglich Zitate der Titel anderer Artikel aus demselben Wörterbuch. Wer sie nicht versteht, schlägt nach und begibt sich auf eine längere Reise entlang der Explikationskette und landet am Schluss eventuell wieder am Ausgangspunkt (Zirkularität).

Es gibt mehrere Möglichkeiten, diese Situation in Bezug auf die Unterscheidung von Mikro- und Makrostruktur zu interpretieren:

a) Die impliziten Verweise (Zitate) auf andere Artikel werden zum Wörterbuchartikel hinzugezählt.

b) Die Verweise werden explizit gemacht und als Verflechtungen zwischen der Mikro- und Makrostruktur einer eigenen Systematik unterworfen.

c) Die Verweise werden überhaupt vermieden, wodurch die Scheidung in Mikro- und Makrostruktur strikt wird.

- Die Interpretation a) ist die einfachste, aber sie ist für die Bedeutungsrepräsentation in maschinenlesbaren Wörterbüchern wenig geeignet, weil dann jeweils für eine Bedeutung die ganze Summe der Explikationsketten gebildet werden müsste, was entweder zu Zirkularität führt oder eine zu lange Kette ergibt. Für den menschlichen Wörterbuchbenutzer genügt diese einfache Bedeutungserklärung mittels impliziten Verweisen, weil beim Menschen die Mehrzahl der Wörter in der Paraphrase als bekannt vorausgesetzt werden können.

- Die Interpretation b) wird auch für maschinenlesbare Wörterbücher sinnvoll, wenn die Verweistypen im Wörterbuch einheitlich sind und wenn die so entstehenden Explikationsketten kurz gehalten werden. An Verweistypen kommen z.B. in Frage:

synonym

hyponym (isa-a)

teil-von (part-of)

besteht aus (set-of)

u.a.

Diese Verweistypen sind bei weitem nicht die einzigen. Grössere Inventare derartiger Verweistypen finden sich in:

- P. Kuhn, der Grundwortschatz (Kühn79) p.82ff

- P.R. Lutzeier, Die semantische Struktur des Lexikons (Lutzeier85) p.106ff

- E. Wüster, Einführung in die Allgemeine Terminologielehre und Terminologische Lexikographie. (Wüster79), p.8ff [**17]

Wenn ein Wörterbuch ein differenziertes Inventar derartiger Verweistypen enthält, kann von einem System von Verflechtungen gesprochen werden, welches neben der Mikro- und der Makrostruktur als dritter konstitutiver Teil des Wörterbuchs angesehen werden muss. Dadurch wird insbesondere der Begriff des Wörterbuchartikels so abgeändert, dass ein Nachschlagen im Diktionär nicht so sehr im Heraussuchen eines "Artikels" besteht, sondern in der Zusammenstellung einer Kette aus mehreren verknüpften Einträgen. Auf diesem Prinzip basiert ein Thesaurus.

- Interpretation c) ist sehr schwierig zu realisieren, denn um die Autonomie jedes Eintrags zu erreichen, müsste jede Bedeutungserklärung zu jedem Eintrag vollständig und ohne Rekurs auf andere Einträge sein.

Diese Möglichkeit besteht, wenn es eine kontrollierte Metasprache gibt in der diese Explikationen vorgenommen werden können. In traditionellen einsprachigen Bedeutungswörterbüchern für den Fremdsprachenunterricht gibt es Ansätze in diese Richtung: Im Explikationsteil wird nur ein restringiertes Vokabular verwendet. Weitergehende Ansätze bauen auf semantischen Primitiva auf, mit deren Hilfe eine Explikationssprache konstruiert wird. Das Set der Primitiva und dessen Organisation als Metasprache kann der Makrostruktur des Wörterbuchs zugerechnet werden, weil die in der Explikationssprache geltenden Regeln für das ganze Wörterbuch gelten.

1.3.1 Analysierbare Teile des Lexikons.

Die drei wesentlichen Teile eines Lexikons sind gemäss den eben gemachten Interpretationen die Makrostruktur, die Mikrostruktur und die Verflechtungsstruktur. Die folgende Aufzählung der Komponenten und Aspekte dieser drei Teile basiert auf den Abschnitten 1.1 und 1.2. und ist deshalb im Sinne einer Zusammenfassung entsprechend knapp formuliert.

a) Makrostruktur:

Diese Struktur kann im Rahmen des Rasters der 7 Typen (nach 1.1.1) als

- Wortliste,
- Folge zweiteiliger Abfolgeschemata,
- Folge variabler Abfolgeschemata,
- Tabellen,
- Baumstrukturen,
- Syntaktisch organisierte Strukturen und
- Wissensrepräsentationsstrukturen

beschrieben werden.

Zu jedem der sieben Typen lassen sich entsprechende Implementationstechniken finden (1.1.2) und für jeden der Typen gelten die in (1.1.3) erläuterten Adäquatheitskriterien wie:

- formale Adäquatheit,
- wissenstheoretische Adäquatheit und
- linguistische Adäquatheit

b) Mikrostruktur:

Eckpfeiler der Beschreibung der Mikrostruktur ist der

- eindeutige (1.2.1),
- disambiguierte (1.2.1) und

-informationslinguistisch fundierte (1.2.3)

Lemmabegriff. Das Lemma wird dabei als eine durch den Lemmatisierungsprozess zu erzeugende Einheit verstanden, welche potentiell lexikalisches Wissen repräsentiert und zur Produktion und Selektion von aktuellem Text- oder Satzwissen eingesetzt werden kann. Dieser Lemmabegriff ist das konkrete Pendant zu dem in (1.1.1) abstrakt postulierten Begriff "Abfolgeschema".

c) Verflechtungsstruktur:

Zur Verflechtungsstruktur gehören:

- implizite Verweise zwischen den Lemmata,
- explizite Verweise zwischen den Lemmata und
- Verbindungen zwischen Lemmata über globale Strukturen und Wertebereiche.

Besonders interessant sind nicht-explizite Verweise. Dazu gehören die impliziten Verweise, wie z.B. vorausgesetzte Kenntnis von Paraphrasewörtern, die selber auch Lemma sind, und die Beziehungen zwischen Einträgen mit Erwartungsrahmen und Einträgen, die in die Leerstellen dieser Rahmen passen.

Man kann diese Arten der nicht-expliziten Verweise als Verbindungen zwischen Lemmata über globale Strukturen und Wertebereiche auffassen. Das lässt sich am Beispiel einer Anfrage an ein datenbankmässig organisiertes Wörterbuch illustrieren:

In einem maschinenlesbaren Wörterbuch können Anfragen an das Lexikon so formuliert werden, dass nicht nur ein "Stichwort" die Antwort bildet (wie das bei gedruckten Wörterbüchern den Normalfall darstellt), sondern dass eine Menge von Lemmata geliefert wird, die das Suchprofil erfüllen. Eine derartige Anfrage wäre beispielsweise

"alle Verben mit zwei obligatorischen Aktanten und dem Kasus Agens=hum".

Falls derartige Lexikoneinträge codiert sind, wird das System eine ganze Menge von Lemmata liefern, welche das Suchprofil erfüllen.

Worin besteht nun die Verflechtung?

Sie besteht bei diesem Beispiel in der inhaltlichen Übereinstimmung der als Antwort generierten Verbeinträge hinsichtlich der Aktantenzahl und des Kasusarguments. Derartige Verflechtungen zwischen den Lemmata basieren auf den den Verben gemeinsamen Strukturen Valenzrahmen und Kasusrahmen sowie auf den für diese Rahmen definierten endlichen Mengen von plausiblen Werten für die Aktanten bzw. Argumente (Wertebereiche). Beim Design eines maschinenlesbaren Wörterbuchs ist besonders auf diese Art der Verflechtung zu achten. Das für das Design der globalen Strukturen und deren Wertebereiche verwendete Metawissen entscheidet über die nicht-expliziten Verflechtungsmöglichkeiten zwischen den Lemmata.

Einheitlichkeit, Systematik und Konsistenz der verwendeten Strukturen und Bereiche entscheiden über die Zuverlässigkeit der in der Abfrage generierten Verbindungen. Erst in der Abfrage werden nicht-explizite Verbindungen zu echten Verflechtungen im herkömmlichen lexikographischen Sinn. Bei maschinenlesbaren Wörterbüchern ist der Aspekt der dynamischen Verflechtung von

zentraler Bedeutung. (Falls das Lexikon in einer Datenbank implementiert ist, sollte allerdings nur der unbedeutendere Teil dynamischer Verflechtungen dem eingebauten DBMS anvertraut werden. Die voraussehbar häufig benötigten Verbindungen sollten im Sinne eines "dedicated System" vorprogrammiert werden).

1.3.2 Knowledge Engineering.

Der Term Knowledge Engineering wird in der Artificial Intelligence verwendet, um die Aktivität zu bezeichnen, Wissen (Fachwissen, Expertenwissen, aber auch allgemeines "Weltwissen") so aufzubereiten, dass es maschinell verarbeitet werden kann. Der Entwurf und die Implementation eines maschinenlesbaren Wörterbuchs fallen auch in dieses Gebiet, denn es geht auch dabei um die Repräsentation von Wissen.

Für die unter (1.3.1) aufgezählten Bauteile des Lexikons

- Makrostruktur,
- Mikrostruktur und
- Verflechtungen

gelten beim Knowledge Engineering die in (1.1.3) genannten Adäquatheitskriterien der Repräsentation. Konkret heisst dies, dass für die Makro-, die Mikro- und die Verflechtungsstruktur je

- formale,
- wissenstheoretische und
- linguistische

Bedingungen eingehalten werden müssen.

Am Beispiel der Verflechtungsstruktur bedeutet dies:

- sie bildet ein formales System mit definierten Elementen und Beziehungen (formal-syntaktischer Aspekt),
- sie gliedert sich in die von Brachman postulierten Repräsentationslevels (wissenstheoretischer Aspekt) und
- sie repräsentiert linguistisches Wissen auf den von Winograd vorgeschlagenen Ebenen (Aspekt der Sprachebenen).

Auf den letzten Punkt fällt das Hauptgewicht, denn Knowledge Engineering heisst in erster Linie "Darstellung von Expertenwissen" und Expertenwissen heisst in diesem Fall linguistisches, lexikographisches und lexikologisches Wissen.

Knowledge Engineering ist eng mit "Knowledge Akquisition" verbunden, d.h. mit dem Erwerb von Wissen, in unserem Fall von lexikographischem Wissen. Für den Wissensakquisition betreibenden Knowledge Engineer heisst die zentrale Frage: Welches sind die relevanten Wissensquellen?

Als Wissensquellen für lexikalische Wissensbasen kommen vor allem in Frage:

- Existierende Wörterbücher,
- Textkorpora,
- lexikographische Arbeiten und
- Grammatiken.

Methoden der Akquisition sind u.a.:

- Abbildung bestehender Wörterbücher in implementierbare Strukturen.
- Extraktion formalisierbarer Einheiten durch Vergleich und Abstraktion bestehender Wörterbücher.
- Lemmatisierung über Texten mit dem Ziel der Lemmagewinnung.
- Beobachtung der Lexikographen bei ihrer Arbeit.

Wieweit diese damit einverstanden sind sog. "speak aloud" Protokolle abzugeben, bleibe dahingestellt.

Besonders wichtig ist in der Praxis die Akquisition von linguistischem Detailwissen z.B. auf den Gebieten der Wortarten und der Ambiguitätsarten, welches auf Sprachkompetenz und linguistischer Erfahrung beruht.

1.3.3 Lexikographische Adäquatheit.

Die lexikographische Adäquatheit der Repräsentation kann theoretisch kaum vollständig abgeleitet werden. Die bereits abgehandelten formalen, wissenstheoretischen und linguistischen Adäquatheitskriterien (in: 1.1.3 und 1.2.3) müssen genügen, denn letztlich beweist Theorie ihre Adäquatheit nur in der Praxis.

Bei maschinenlesbaren Wörterbüchern gibt es eine Validierung durch die Praxis in erster Linie in der Implementation.

Von diesen praktischen Prüfsteinen lexikographischer Adäquatheit handelt Teil 2 vorliegender Arbeit, indem die zu 6 implementierten Lexikonsystemen erhobenen empirischen Daten vorgestellt und kommentiert werden.

TEIL 2 EMPIRISCHE ERKENNTNISSE AUS EINER UMFRAGE.

Um zu empirischen Erkenntnissen über maschinenlesbare Wörterbücher zu gelangen, wurde die Methode einer Umfrage gewählt. Eine erste Motivation für eine Umfrage über implementierte maschinenlesbare Wörterbücher ergibt sich aus dem Schwerpunkt "Semantik" dieser Arbeit. In der linguistischen Forschung wurde bis heute (1986) keine Einigung darüber erzielt, was als das "Grundprädikat" der Semantik bezeichnet werden könnte. Diese Ansicht wird von anerkannten Kennern vertreten. Stellvertretend für andere ähnliche Formulierungen sei folgendes negatives Fazit zitiert:

"Seit Chomsky besteht eine quasi-universelle Einigung darüber, was das Grundprädikat der Syntax sein muss. Dem ist in der Semantik leider nicht so. Man ist sich nicht nur über die Bestimmung des Grundprädikats der Semantik nicht einig, es gibt auch so gut wie keinen Konsens über die Form und Zielsetzung einer semantischen Theorie. Diese hat viele Gründe, die wir hier nicht erörtern wollen."

(Guenthner/Lehmann86, p.164)

Angesichts dieser Sachlage scheint es sinnvoll zu sein, empirisch vorzugehen und zu dokumentieren, wie "Semantik" in real existierenden Lexika Sprachdaten verarbeitender Systeme konzipiert und implementiert ist.

2.1 * METHODE UND ZIEL DER UMFRAGE.

Das Ziel der Umfrage besteht deshalb u.a. darin, Aufschluss darüber zu erhalten, wie semantische Information in maschinenlesbaren Wörterbüchern implementiert wird. Da auch bei empirischer Forschung theoretische Überlegungen impliziert sind, ist es von Vorteil, solche schon zu Beginn zu explizieren. Guenthner und Lehmann unternehmen im eben zitierten Artikel einen theoretischen Versuch, die "zwei bekanntesten Ansätze" der heute praktizierten Auffassungen semantischer Repräsentation generell zu taxieren.

"Es ist vielleicht sinnvoll, die zwei bekanntesten Auffassungen zu Form und Zielsetzung einer Semantiktheorie kurz zu kennzeichnen und auf eine mögliche Vereinigung dieser beiden Auffassungen hinzuweisen. Linguistische Semantiktheorien sowie die meisten Ansätze im Rahmen der maschinellen Verarbeitung natürlicher Sprache und in der künstlichen Intelligenz neigen seit jeher zu einer repräsentationellen Form der Semantik. Aufgabe der Semantik ist hier die Übersetzung von syntaktischen Analysen in eine semantische Repräsentationssprache. Die Vorschläge reichen von Merkmalskonstrukten (im Sinne der generativen Grammatik) bis zu

verschiedenen Netz- und Framekonstrukten, wie sie von Minsky, Schank, Hendrix und vielen anderen vorgeschlagen wurden. Die Bedeutung eines Satzes wird somit in einer künstlichen bzw. formalen Sprache dargestellt.

Im Gegensatz zu den repräsentationellen Ansätzen versucht man in der logischen (oder modelltheoretischen) Semantik, die Bedeutung eines Satzes durch seine Wahrheitsbedingungen zu erfassen. In den meisten Arbeiten (vor allem in solchen, die auf der sogenannten Montague-Grammatik in irgendeiner Weise aufbauen) werden die Wahrheitsbedingungen in bezug auf Modelle (oder 'mögliche Welten') angegeben. In gewisser Weise beinhaltet eine modelltheoretische Semantik auch einen repräsentationellen Anspruch, denn wenn die Wahrheitsbedingungen für einen Satz der natürlichen Sprache durch die Semantik für dessen Übersetzung z.B. in die Prädikatenlogik gegeben werden, kann man die Übersetzungen selber als semantische Repräsentation betrachten."

(Guenthner/Lehmann86, p.164)

Die Ähnlichkeit der beiden Vorstellungen besteht nach Guenthner/Lehmann darin, dass sowohl die diversen repräsentationellen Konstrukte als auch prädikatenlogische Modelle das Resultat eines Übersetzungsprozesses sind, der sich unter Umständen aus mehreren Stufen von Transformationen zusammensetzt. Der Gegensatz "repräsentationell" und "modelltheoretisch" bezieht sich deshalb hauptsächlich auf den Unterschied in der Art des Übersetzungsprozesses und auf die verwendeten Übersetzungsregeln.

In "Semantikkonzepte in der Künstlichen Intelligenz" weist C. Schwind auf diese Unterschiede hin:

"Die Übersetzungsregeln sind Abbildungen von syntaktischen auf semantische Strukturen. Sie können in die Syntaxregeln eingebaut sein. In diesem Fall wird der Satz schrittweise während der syntaktischen Analyse und parallel zu dieser in die semantische Repräsentation übersetzt."

(Schwind85, p.98)

Eine alternative Technik ist ein mehrstufiger (serieller) Übersetzungsprozess, bei dem Syntax und Grammatik getrennt werden.

"Die Übersetzungsregeln können jedoch auch unabhängig von den Syntaxregeln direkt als Abbildungen von syntaktischen auf semantische Strukturen formuliert sein. In diesem Fall wird der Satz zunächst vollständig syntaktisch analysiert und die Übersetzungsregeln werden anschliessend angewendet und operieren nur auf syntaktischen Strukturen."

(Schwind85, p.98)

Bei dem von Guenthner und Lehmann beschriebenen Verfahren im juristischen Expertensystem LEX (Forschungszentrum IBM Heidelberg) lässt sich der Übersetzungsprozess folgendermassen beschreiben:

"Wir erzeugen Diskursrepräsentationsstrukturen auf der Grundlage des USL Systems [**1] (s. auch Guenthner/Lehmann (1984) für eine genauere Beschreibung). Das bedeutet, dass im Anschluss an die Syntaxanalyse und die Erzeugung von Zwischenstrukturen aus diesen statt SQL-Ausdrücken DRSen [**2] generiert werden (zunächst ohne die Auflösung kontextueller Referenzen). Erst in einem weiteren Schritt werden diese erkannt, wobei morphologische, syntaktische,

semantische und pragmatische Kriterien herangezogen werden (s. Guenther / Lehmann (1983) für eine detaillierte Diskussion der Behandlung von Pronominalisierung)." [**3]
(Lehmann85, p.95)

Das Resultat des Übersetzungsprozesses hängt von den in den Zwischentransformationen verwendeten Regeln ab. Lehmann und Guenther (wie übrigens auch C. Schwind) bevorzugen aus Gründen besserer Manipulationsmöglichkeiten solche Transformationen, die im Resultat eine logische Semantikrepräsentation (d.h. in der Terminologie von Guenther/Lehmann die modelltheoretische Semantik) erzeugen:

"Die Vorteile des modelltheoretischen Ansatzes sowohl für die theoretische Linguistik als auch für die Verarbeitung natürlicher Sprachen liegen auf der Hand. Gerade beim Aufbau von natürlichsprachlichen Datenbank-Abfragesystemen spielt die Möglichkeit, die Semantik von Abfragen, sowie von Updates prädikatenlogisch zu erfassen, eine grosse Rolle. Relationale Datenbanken sind im Grunde Modelle im Sinne der modelltheoretischen Semantik (wenigstens in den meisten Anwendungen, in denen die sog. 'closed-world assumption' gemacht wird), und die Interpretation einer Abfrage entspricht nichts anderem als der operationalisierten Anwendung von Wahrheitsbedingungen. Genausogut kann man aber auch Datenbanken und die Abfrage deduktiv anlegen."
(Guenther/Lehmann86, p.164)

Diese Manipulierbarkeit logischer Semantikstrukture in Gestalt der Kombination relationaler Datenbanktechnik mit Deduktionsmaschinen entspricht der Konzeption DB-basierter Expertensysteme, wie sie z.B in (Appelrath85), p.28ff., dargestellt wird, also m.a.W. in der Kombination SQL-ähnlicher Kalküle mit PROLOG-ähnlichen. Allerdings ist diese Parallele mit Vorsicht zu geniessen, denn in dem von Guenther und Lehmann beschriebenen "juristischen Expertensystem auf der Grundlage von Linguistik und Logik" (LEX) der IBM ist der betriebene Aufwand schon noch etwas grösser. [**4]

Die Intention hinter solch aufwendiger logischer Repräsentation ist in der effizienten und konsistenzhaltenden Manipulierbarkeit logischer Semantikmodelle bei Inferenzprozessen zu sehen. Im Kern wird dieser Gedanke von Guenther und Lehmann so zusammengefasst:

"Die logisch-orientierten Ansätze haben darüber hinaus den gewichtigen Vorteil, dass es für die Repräsentationssprache (diese ist in den meisten Fällen die Prädikatenlogik der 1. Stufe oder eine notationelle Variante davon, etwa eine mehrsortige Sprache) zwei verschiedene, wenn auch im Endeffekt äquivalente Manipulationsmöglichkeiten gibt: Einerseits kann man für die Ausdrücke der Sprache semantische Regeln und Wahrheitsbedingungen präzise formulieren (die modelltheoretische Semantik), andererseits können prädikatenlogische Formeln auch deduktiv manipuliert werden. Diese Manipulationsmöglichkeiten sind ebenso wichtig wie die Repräsentationen selbst, denn es geht in der Semantik nicht nur darum, die Bedeutung von Sätzen in irgendeiner Form zu erfassen, man muss auch Repräsentationen bzw. Bedeutungen in Beziehung setzen können, (z.B. um zu klären, ob aus einer gegebenen Repräsentation oder Menge Repräsentationen eine andere folgt oder nicht)."

(Guenthner/Lehmann86, p.164)

*

Durch die im folgenden präsentierten Projekte und Wörterbücher wird diese klare Stellungnahme für ein logisches Semantikmodell empirisch relativiert werden. Damit ergibt sich neben der eingangs erwähnten Suche nach dem "Grundprädikat" in der Semantik eine weitere Motivation für die durchgeführte empirische Untersuchung.

2.1.1 Auswahl der Projekte.

Die Auswahl der untersuchten Wörterbücher wird durch folgende Figur illustriert:

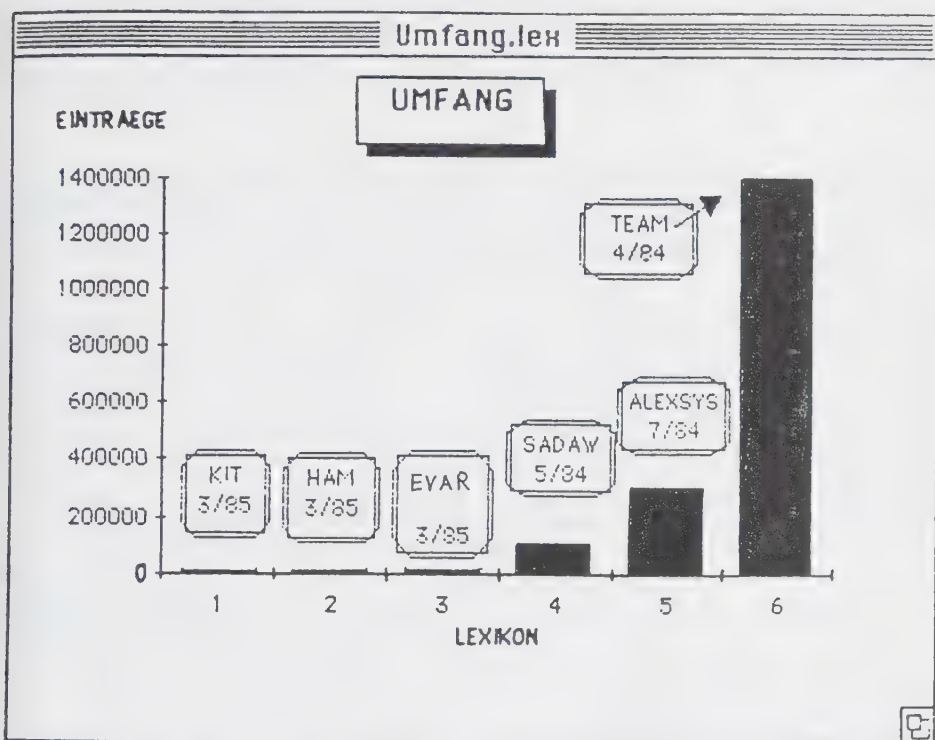


Figure 5. 'Umfang ausgewählter Lexika'

Die genauen Grössenwerte wurden zum Zeitpunkt der Untersuchung wie folgt angegeben:

HAM	KIT	EVAR	SFB100	ALEXSYS	TEAM
ca 1500 GF	einige 100 SF	ca 5000 VF	ca 110'000 GF	ca 300'000 VF/GF	ca 1,5-2 Mio VF

(SF=Stammformen, GF=Grundformen, VF=Vollformen)

Diese Grössenangaben wurden als Antwort auf die Frage Nr. 13 des Fragebogens angegeben, welche lautet:
 "Wie gross ist Ihr Wörterbuch? (Anzahl Einträge/ Anzahl Grundformen/ Anzahl abgeleitete Formen/ andere Grössenmerkmale)".

Die gegebenen Antworten wurden auch mit Frage Nr. 1 des Fragebogens abgeglichen, welche lautet:

"Verwenden Sie mehrere Typen von Lexika? (Vollformenwörterbuch/ Grundformenwörterbuch/ Stammwörterbuch/ andere)".

Die Zeitpunkte der Erhebung und damit die Gültigkeit der Angaben sind:

HAM	KIT	EVAR	SFB100	ALEXSYS	TEAM
3/85	3/85	3/85	5/84	7/84	4/84

Die Angaben auf den beantworteten Fragebogen wurden von folgenden Personen verfasst: [**5]

HAM	KIT	EVAR	SFB100	ALEXSYS	TEAM
Dr. W. Hoeppner	Dr. M. Eimermacher	Dr. A. Brietzmann	Prof. Dr. H.-D. Maas	Dr. J. Brustkern	Dr. E. Hanis

Die für die Projekte verwendeten Akronyme haben folgende Bedeutung:

HAM: Hamburger Anwendungsorientiertes Natürlichsprachliches System.

KIT: Kognitive Verfahren zur Informations-Extraktion und Zusammenfassung aus Texten.

Eine neuere Lesart des Akronyms KIT bedeutet:

Künstliche Intelligenz und Textverstehen. (Rollinger85, p27)

EVAR: Erkennen-Verstehen-Antworten-Rückfragen.

SADAW: Saarbrücker Deutsches Analyse-Wörterbuch.

ALEXSYS: ALLgemeines LEXikonSYstem.

TEAM: Terminology Evaluation And Management.

*

Die Auswahl gerade dieser sechs Lexikonsysteme hat mehrere Gründe:

1. KI-Forschung, Linguistische Forschung und praktische Nutzung sind repräsentiert.

2. Kleinere und umfangreichere Systeme sind vertreten.

3. Die Systeme sind mehrheitlich monolingual und auf das Deutsche bezogen.

4. Die Auswahl bleibt überschaubar.

5. Ehrlicherweise muss auch gesagt werden, dass die Vollständigkeit der Antworten auf den zurückerhaltenen Fragebogen ein nicht zu unterschätzendes Auswahlkriterium bildete.

2.1.2 Die Formulierung der Fragen.

Um die Art der gestellten Fragen zu charakterisieren, mögen zwei Beispiele dienen:

- In welcher Programmiersprache operieren Sie über dem Wörterbuch?

HAM	KIT	EVAR	SFB100	ALEXSYS	TEAM
UCI-LISP	PROLOG	SIMULA neu: LISP	FORTRAN Ling. Untprg	PL1	ASSEMB- LER

- "Welchen Zwecken dient das Lexikon? (Lemmatisierung/ Syntax/ Semantik/ M.Ü./ I.u.D./ K.I./ andere)".

(Die in der Frage benutzten Abkürzungen bedeuten:

M.Ü. = Maschinelle Übersetzung

I.u.D. = Information und Dokumentation

K.I. = Künstliche Intelligenz)

HAM	KIT	EVAR	SFB100	ALEXSYS	TEAM
LEM. SYN. K.I.	SYN. SEM. K.I.	SYN. SEM. K.I. PHON.	SYN. SEM. M.Ü. DOKU. (CTX)	SYN. SEM. M.Ü. DOKU. K.I. ALLG WORDTB	M.Ü. TERM. WÖRTER BUCH HERSTLG

Wie man sieht, wurden die Fragen teils im multiple-choice Verfahren gestellt, teils offen gehalten.

Abgesehen vom Nutzen der ausgefüllten Fragebogen bei der Charakterisierung der einzelnen Lexika, erfüllt der Fragebogen noch einen anderen Zweck: Er kann als Basis für einen Vergleich der 6 doch recht unterschiedlichen Lexikonsysteme dienen. Dies hängt direkt mit der Formulierung der Fragen zusammen, welche eine kompakte synoptische Darstellung der Antworten zulässt. (Vergl. TEIL 3)

2.1.3 Echo, Präsentation, Auswertung.

Das Echo auf den Fragebogen war in den meisten Fällen ausserordentlich gut, indem den meisten zurückgeschickten Bogen reichlich Arbeitsberichte und Literaturhinweise beigelegt waren.

Präsentiert wird das Antwortmaterial in drei Stufen:

1. In unveränderter Form und nach Projekten geordnet (Kapitel 2.2.1 bis 2.2.6). Auf dieser 1. Stufe wird auch auf den Projektkontext hingewiesen; ebenso auf fehlende Antworten und eventuell auf andere Informationsquellen, die die Lücken ausfüllen könnten.

2. In synoptischer Form nach Themenbereichen geordnet. (TEIL 3) . Die dabei entstehenden kleinen Tabellen werden kommentiert.

In einem 3. Schritt wird beim Vergleich der Antworten das Thema Semantik eingehender kommentiert (Kapitel 3.2).

Der synoptische Vergleich nach Themenbereichen und die Kommentierung der Antworten in Bezug auf das Thema Semantik stellen bereits einen Teil der Auswertung dar.

Weitere Aspekte der Auswertung kommen im Kapitel 3.3 zur Sprache, wo auch der Erkenntniswert der Umfrage abgeschätzt werden soll.

2.2 * DIE ANTWORTEN NACH PROJEKTEN GEORDNET.

In (2.2.1) bis (2.2.6) werden die Antworten nach Projekten originalgetreu wiedergegeben.

In TEIL 3 werden die Daten verglichen; im Interesse besserer Lesbarkeit wurden Redundanzen z.T. beibehalten.

Die den Antworten vorgeordneten Projektbeschreibungen sind kompakt gehalten und sollen dazu dienen, das Lexikon des jeweiligen Projekts im Systemzusammenhang zu sehen. Die auf die Antworten folgenden Ergänzungen und Hinweise dienen primär dazu, weitere Informationsquellen zu nennen.

2.2.1 Das Lexikon des Systems HAM-ANS.

Das Lexikon wird im Projekt als "lexikalische Wissensquelle" bezeichnet. Von diesen Quellen gibt es mehrere: Wortlexikon,

Syntagmenlexikon, Kasusrahmenlexikon, Stammformenlexikon und Flexionsinformationen.
(vergl. Hoepfner83b)

Das Projekt.

Die Projektbeschreibung basiert auf (Bergmann82), (Hoepfner83a,b), (Busemann84) und (Hoepfner82).

HAM-ANS, Hamburger Anwendungsorientiertes Natürlichsprachliches System, ist das Nachfolgesystem von HAM-RPM. Beide Systeme sind Projekte der Universität Hamburg, Forschungsstelle für Informationswissenschaft und Künstliche Intelligenz. RPM, das Redepartnermodell, wurde als Simulationsmodell konzipiert, welches das Redeverhalten eines Hotelverwalters bei der Zimmervermietung imitiert (v.Hahn76, v.Hahn80). In HAM-ANS wird die Beschränkung auf einen geschlossenen Diskursbereich zugunsten mehrerer Modelle aufgehoben. In HAM-ANS wird die Entwicklung eines geschlossenen Simulationsmodells zu einer natürlichsprachlichen Schnittstelle für unterschiedliche Applikationen vollzogen. Die Applikationen wurden mit Absicht so gewählt, das sich im Verkehr mit dem Benutzer über die NL-Schnittstelle unterschiedliche Dialogtypen ergeben.

1. In der aus RPM übernommenen Dialogsituation der Hotelzimmer-Vermietung wird der Dialog von Seiten des Systems durch das Verkaufsinteresse des "Hotelverwalters" zielstrebig in eine Richtung vorangetrieben. Das zu diesem Dialogtyp passende Systemverhalten kann als "interest-based" charakterisiert werden. Der Diskursbereich dieser Applikation lässt sich mit "Alltagswissen" umschreiben, wobei aber zwischen dem Systemwissen und dem im Dialogpartnermodell vorausgesetzten Benutzerwissen differenziert werden muss. Das Systemwissen ist insofern Expertenwissen und nicht Alltagswissen, als zum Systemwissen Verkaufsstrategiewissen, Hotelfachwissen, Dialogsteuerungswissen und sprachliches Wissen gehören. Im Benutzermodell wird dem Benutzer dagegen lediglich sprachliches Wissen und Alltagswissen über Hotels attestiert. Deshalb wird dieser Dialog vom System dominiert. Die Dialogführung basiert auf einem Expertensystem, welches über die oben genannten Wissensquellen verfügt (implementiert in Lisp und Fuzzy). An diesem 1. Dialogtyp wurden im Projekt HAM-ANS vor allem die folgenden linguistischen Gebiete erforscht:

- a) Explizite Dialogpartner-Modellierung.
- b) Dialogstrategien.
- c) Erzeugung von interessegesteuerten Antworten.

Es ist naheliegend, dass die erzielten Forschungsergebnisse nicht nur im Bereich der Hotel-Reservierung angewendet werden können.

2. Der zweite in HAM-ANS realisierte Dialogtyp unterscheidet sich sowohl hinsichtlich des Diskursbereichs als auch in Bezug auf die Dialogstrategie vom RPM-Verkaufsdialo. Die Applikation besteht in einem visuellen Beobachtungssystem. Beobachtet wird der Straßenverkehr auf einer Kreuzung. Die visuelle Information wird durch das Bildsequenzanalyse-System NAOS/MORIO in eine Folge von zeitlich determinierten geometrischen Beschreibungen transformiert. [**6]

Vom Beobachter wird angenommen, dass er die beobachtete Szene (Strassenkreuzung) zwar kennt, aber nicht sieht. Lediglich über die NL-Schnittstelle kann er sich detaillierte Informationen über die aktuelle Verkehrslage auf der Kreuzung verschaffen.

Der zu dieser Situation passende Dialogtyp kann als "cooperative" bezeichnet werden, weil das System die Informationswünsche des Benutzers in idealer Weise befriedigen soll. Im Unterschied zum Verkaufsdialog soll dem Benutzer kein geschöntes Bild der Realität vermittelt werden. Bei der Antwortgenerierung wird deshalb auf tendenziöse Ausschmückung der Situationsbeschreibung (etwa mit positiv wertenden Adjektiven) verzichtet.

Der Diskursbereich dieser 2. Applikation lässt sich zwar ebenfalls mit "Alltagswissen" umschreiben, aber im Unterschied zur Hotelreservierung besteht die Wissensdifferenz zwischen System und Benutzer beim Verkehrsbeobachtungssystem eher in quantitativer als in qualitativer Hinsicht. Der Benutzer kennt die aktuelle Situation nicht so genau wie das System, aber das Systemverhalten ist primär darauf angelegt, diese Differenz zu verringern, nicht sie kommerziell auszunutzen.

An diesem Dialogtyp wurden im Projekt HAM-ANS u.a. folgende Gebiete erforscht:

- a) Semantische Repräsentierung von Raum und Zeit.
- b) Referenzsemantik von Bewegungsverben.
- c) Referenzsemantik sowohl im Zusammenhang mit der Generierung von eindeutigen Referenzen in den Antworten als auch im Hinblick auf die eindeutige Auflösung der Anaphern des Benutzers.

Auch in Bezug auf diese Forschungen ist anzunehmen, dass die Resultate nicht nur zur Überwachung von Strassenkreuzungen eingesetzt werden können.

3. Der dritte in HAM-ANS realisierte Dialogtyp unterscheidet sich von den beiden genannten durch einen sehr speziellen Diskursbereich. Ansonsten ist dieser Dialogtyp am Muster eines natürlichsprachlichen Datenbankinterface's orientiert.

"In the database application HAM-ANS provides marine scientists with NL access to a fishery database, which contains data collected during international expeditions in the South Atlantic and the Indian Ocean. The complex structured data sets, which consist of cruise data concerning the vessels voyages, krill biology data, relevant net hauls and oceanographic data, are stored in a relational database with a Pascal/R data base management system. Pascal/R is an extension of the well-known programming language PASCAL, having the data structure relation and a general content-based selection mechanism for relations"
(Hoepfner83a, p.588)

Die Forschungsschwerpunkte für diese dritte Applikation werden in (Hoepfner83a) wie folgt angegeben:

- a) "Coping with mass data".
- b) "Representation and acquisition of meta-knowledge about databases."
- c) "Use of dialog coherence for query optimization."

Als Anwendungsgebiet für Datenbanken mit natürlichsprachlicher Schnittstelle kommt jedes Fachgebiet in Frage.

Die Antworten.

1. FRAGE:

Verwenden Sie mehrere Typen von Lexika? (Vollformenwörterbuch/ Grundformenwörterbuch/ Stammwörterbuch/ andere)

ANTWORT:

Grundformenwörterbuch

Zusatzbemerkung:

s. Arbeitspapier 15

2. FRAGE:

Welchen Zwecken dient das Lexikon? (Lemmatisierung/ Syntax/ Semantik/ maschinelle Übersetzung/ Information, Dokumentation/ Künstliche Intelligenz/ andere)

ANTWORT:

Lemmatisierung, Syntax, Künstliche Intelligenz

3. FRAGE:

Welches Wortklassenschema wird zugrundegelegt?

- Wortarten (Tabelle, ev. Abkürzungen).
- Welche Wortarten unterteilen Sie in welche Unterklassen?
- Welche Wortarten fassen Sie zu welchen Oberklassen zusammen?

ANTWORT:

s. Busemann (Rep.27), s.17-19

4. FRAGE:

Welche Informationsfelder werden pro Eintrag codiert? (zusammengehörende Informationsfelder (Codes) als solche markieren) Falls die Einträge nach Wortarten differenziert codiert sind, dies in analogen Tabellen darstellen.

ANTWORT:

s. Arbeitspapier 15, s.1

5. FRAGE:

Welchen Rechner benutzen Sie?

ANTWORT:

6. FRAGE:

Ist Ihr Verfahren fileorientiert oder datenbankorientiert?

ANTWORT:

datenbankorientiert

Frage 7 fällt weg (Vergl. Anhang)

8. FRAGE:

Falls datenbankorientiert:

- welchen DB-typ verwenden Sie? (CODASYL/ Baumhierarchie/ Inverted File/ Relational/ andere)

ANTWORT:

Eigentlich eine selbst entwickelte, in LISP eingebettete DB

- Die nächste Teilfrage entfällt.

TEILFRAGE: -Wie lässt sich die Abspeicherung eines Worteintrags in der Terminologie Ihres Datenbanksystems beschreiben?

ANTWORT:

Eintrag + Feature-value-Paare

TEILFRAGE:

- Welche Möglichkeiten vermissen Sie bei Ihrem DBMS?

ANTWORT:

--

9. FRAGE:

Sind ausser der codierten Lexikoninformation noch andere Informationen codiert? (Affixlisten/ andere)

ANTWORT:

Affixlisten

10. FRAGE:

In welcher Programmiersprache operieren Sie über dem Wörterbuch?

ANTWORT:

UCI-LISP

11. FRAGE:

Mit welchen Zugriffsalgorithmen operieren Sie? (VSAM/ hash/ B-Baum/ andere)

ANTWORT:

hash

12. FRAGE:

Welche Felder können als Zugriffsschlüssel verwendet werden? (nur ein Schlüssel/ alle Felder/ andere)

ANTWORT:

nur ein Schlüssel

13. FRAGE:

Wie gross ist Ihr Wörterbuch? (Anzahl Einträge/ Anzahl Grundformen/ Anzahl abgeleitete Formen/ andere Grössemerkmale)

ANTWORT:

ca 1500 Einträge

14. FRAGE:

Maximale Wortlänge?

ANTWORT:

nicht begrenzt

15. FRAGE: Maximale Anzahl von Informationskategorien (Codes) pro Lexem?

ANTWORT:

ca 7

16. FRAGE:

Besteht die Möglichkeit, Einträge zu verknüpfen? (Verknüpfung Lemma- Formen/ Thesaurusrelationen/ Frames/ andere

ANTWORT:

Verknüpfung Lemma-Formen.
Kasusrahmenlexikon.

17. FRAGE:

Welche LDV-Programme (oder Systeme) benutzen Ihre Wörterbücher?

ANTWORT:

HAM-ANS und verschiedene Versionen der
Oberflächentransformationen auf Symbolics-Rechnern.

18. FRAGE:

existiert eine Schnittstelle zur DIN 2341 (z.B. MATER)? [**7]

ANTWORT:

nein

19. FRAGE:

Welche konventionellen Wörterbücher dienten als Grundlage? (für den Wortschatz/ für die Codierung/ andere Informationen)

ANTWORT:

--

20. FRAGE:

Ist es Ihnen möglich, einen typischen Auszug aus dem Wörterbuch und eine Erklärung des Codes beizulegen?

ANTWORT:

--

21. FRAGE:

Literaturhinweise:

ANTWORT:

HAM-ANS Report 27

Ergänzungen und Hinweise.

Diese Hinweise wurden zusätzlich zu den Antworten auf den Fragebogen notiert.

1. Hinweis:

Projekt am IKP in Bonn. Abschlussbericht:

Deutsche Wörterbücher, Dokumentation, Vergleich, Integration, Tübingen, 1983.

2. Hinweis:

Arbeitstreffen "Spezifikation einer lexikalischen Wissensbasis (Deutsch) für natürlichsprachliche KI-Systeme", das im Januar 1983 in Bonn stattfand. Bericht darüber im KI-Rundbrief.

3. Hinweis: Universität Erlangen, Lehrstuhl Prof. Niemann.

R. Müller-Meernach, Datenbasis für den Lexikonexperten eines Spracherkennungssystems, Studienarbeit, 1981.

Abgesehen von diesen 3 Hinweisen wurden folgende Arbeitsberichte beigelegt:

1. Wolfgang Hoepfner, Lexikalische Wissensquellen im Dialogsystem HAM-ANS. in: Arbeitspapier Nr.15 der Forschungsstelle für Informationswissenschaft und Künstliche Intelligenz, Universität Hamburg, 1983

2. Stephan Busemann, Surface Transformations during the Generation of written German Sentences. In: Report ANS-27 of the Research Unit for Information Science and Artificial Intelligence, Univ. Hamburg 1984.

Die Wortklassenangaben in (Busemann84), auf die in der Antwort auf die Frage Nr. 3 verwiesen wurde, sind aus den beiden nächsten Figuren abzulesen. Das erste Schema repräsentiert die Wortklassen für die Oberflächentransformationen (SUTRA = Surface Transformation), das zweite Schema zeigt die für die Generation verwendeten Wortklassen.

• Verbs, auxiliaries, modal verbs	VRB
• Nouns	NOM
• Attributive adjectives	ADJ
• Definite article (D)	DETDEF
• Indefinite article (EIN), KEIN	DETIND
• Quantifiers and Determiners in article position	DETTQ
• Relative pronouns	RELPRN
• Personal pronouns	PERSPRN
• Possessive pronouns	POSSPRN
• Prepositions	PRP
• Remaining noninflecting stems	XXX

Figure 6. 'SUTRA-Wortklassen (HAM-ANS)'

Die SUTRA-Wortklassen lassen sich in die Wortklassen für die Generation projizieren:

	<u>HAM-ANS</u>	<u>SUTRA</u>
Verb <i>SEIN, HAB, WERD</i>	VRB AUX }	VRB
Noun	NOM	NOM
Attributive adjective Attributive quantifying adjective Attributive perfect participle	ADJ DET PAR }	ADJ
Article and quantifier	DET	{ DETDEF DETIND DETQ
Personal pronoun	PRN	PRN
Possessive pronoun	POSSPRN	POSSPRN
Preposition	PRP	PRP
Predicate adjective Predicate perfect participle Adverb Verb prefix Conjunction Subordinating conjunction Interjection Cardinal number Negation particle	ADJ PAR ADV VRBZ KON NSKON ITJ ZWT NEG }	XXX

Figure 7. 'Forts. Wortklassen (HAM-ANS)'

*

Die nun folgenden Angaben aus (Hoepfner83b) geben Aufschluss über die in HAM-ANS verwendeten heterogenen lexikalischen Wissensquellen.

<u>lexikal. Eigenschaft</u>	<u>Wert</u>	<u>Beispiel</u>
WORTART	Name einer Wortklasse	HAUS (WORTART NOM)
STAMM	Grundform	GLAESER (STAMM GLAS)
SYNONYM	Abbildung eines Wortes auf ein anderes oder mehrere andere	AM (SYNONYM AN DEM)
FLEXION	Flexionsmerkmale	DICH (FLEXION (SG (AKK)))
KASUSRAHMEN	Name eines Kasusrahmenkonstruktors	BREMS (KASUSRAHMEN AGENT-LOCATIVE-3)
GENUS	Teilmenge von (MAS FEM NTR) für Nomen und Pro- nomen	BEIN (GENUS NTR)
DEKLIN	Kennzeichnung von Nomen- klassen für die Generierung	BILD (DEKLIN ER)
KONJUG	Für die Generierung be- nutzte Kennzeichnung von Verbklassen	BIEG (KONJUG (VTYP 4) (VGFC BOG) (VGFD BOEGE) (PAUX HAB) (PPRF GEBOGEN))

Figure 8. 'Relationen im Wortlexikon (HAM-ANS)'

Im System HAM-ANS gibt es ausser dem Wortlexikon noch andere lexikalische Wissensquellen. Diese sind:

1. Das Syntagmenlexikon.

"Ausser durch die Einträge der Wissensquelle Wortlexikon werden durch das Syntagmenlexikon einzelne Wörter im jeweiligen Satzzusammenhang schon während der lexikalischen Analyse verändert. Drei Möglichkeiten stehen hier zur Verfügung: Es können Wörter durch andere ersetzt werden (mittels der Aktion ERSETZEN), mehrere Wörter können zu einem zusammengefasst werden (Aktion KOMPRIMIEREN) oder es kann eine direkte Reaktion auf die Benutzeräusserung erfolgen (Aktion ERWIDERN)."

Ein Beispiel für die Aktion KOMPRIMIEREN wäre:

"(DEFPROP NAEHE ((KOMPRIMIEREN IN DER NAEHE)) IDIOME)"

(Hoepfner83, p.2)

Das Syntagmenlexikon enthielt 1983 "knapp 100 verschiedene Transformationen".

(Hoepfner83, p.2)

2. Das Kasusrahmenlexikon:

"Im Wortlexikon steht für einige Verben unter der Eigenschaft KASUSRAHMEN ein Verweis auf Einträge des Kasusrahmenlexikons."

(Hoepfner83, p.2)

"Zur Zeit [1983 H.H.] enthält das Kasusrahmenlexikon erst 8 Konstrukturen für verschiedene Ausprägungen von Kasusrahmen."

(Hoepfner83, p.2)

Zur Verwendung von Kasusrahmen in HAM-ANS vergl. (Hoepfner82).

3. Flexionssuffixe:

"In HAM-ANS wird ein Stammformen-Lexikon verwendet. Die Lemmatisierung erzeugt nicht nur die Grundform flektierter Wörter durch Identifikation von Suffixen, Umlaut oder Partizipialpräfixen, sondern auch die morphologisch-syntaktischen Merkmale der Ausgangsform"

(Hoepfner83, p.3)

"Dies geschieht über eine Zuordnung von Flexionssuffixen und anderer verfügbarer Information (z.B. Wortklasse, Genus) zu einer Beschreibung ihrer Merkmale (vgl. (Bergmann82))."

(Hoepfner83, p.3)

4. Schreibfehlerkorrektur:

"Für das System zur Schreibfehlerkorrektur ist eine Matrix aus Bit-Listen als interne Repräsentation der Einträge des Wortlexikons Bestandteil der lexikalischen Wissensquellen von HAM-ANS".

(Hoepfner83, p.5)

5. Semantische Netze:

"Eigennamen werden z.B. über Netzrelationen identifiziert und erhalten die morphologischen Merkmale ihrer Oberbegriffe. Insbesondere für den Wissenserwerb wird eine Verzahnung zwischen lexikalischen und semantischen Wissensquellen notwendig"

(Hoepfner83, p.5)

6. Die Repräsentationssprachen SURF und DEEP:

"Eine weitere Wissensquelle, die das lexikalische Klassifikationsschema ergänzt, ist die Definition der semantischen

Repräsentationssprachen SURF und DEEP, in der bestimmte Wörter, z.B. Quantoren und Partikeln, in verschiedene Klassen subkategorisiert werden. Durch diese Einteilung, die eher an Problemen der Auswertung und der Referenz orientiert ist, wird eine Re-interpretation geleistet, die sich auch für andere Zwecke einsetzen liesse, etwa für das Abweichen von einer sonst gültigen Klassifikation in bestimmten Verarbeitungsumgebungen." (Hoepfner83, p.5)

2.2.2 Das Lexikon des Systems KIT.

Die Lexikoneinträge werden als "Wortexperten" bezeichnet. Es handelt sich dabei um Stammformeneinträge mit assoziierten Regeln.

Das Projekt.

Die Projektstruktur von KIT wird in (Rollinger85) in kompakter Form beschrieben, sodass es angemessen ist, diese Beschreibung im Original zu übernehmen:

"An der Technischen Universität Berlin, am Institut für Angewandte Informatik, werden in der Lehr- und Forschungseinheit "Computergestützte Informationssysteme (CIS)" unter der Leitung von Prof. Dr. H.-J. Schneider unter der Bezeichnung "Projektgruppe KIT" derzeit vier Forschungsprojekte aus dem Bereich der Künstlichen Intelligenz durchgeführt. Die Abkürzung KIT geht zurück auf das Projekt "Kognitive Verfahren zur Informationsextraktion und Zusammenfassung aus Texten" und bedeutet seit nunmehr über einem Jahr "Künstliche Intelligenz und Textverstehen". Diese im nachfolgenden kurz beschriebenen Projekte sind hervorgegangen aus einer Reihe von Vorläuferprojekten, die seit 1975 an dieser Einheit durchgeführt wurden und von dem BMFT über die Gesellschaft für Information und Dokumentation (GID) finanziert wurden. Dabei handelte es sich um die Projekte

"Automatische Erstellung semantischer Netze" (1975-1978 Projektleiter V. Cherniavsky; 1978-1981 Projektleiter Ch. Habel),

"Kognitive Verfahren zur Informationsextraktion und Zusammenfassung aus Texten" (1981-1983, Projektleiter C.-R. Rollinger) und

"Textverstehen: Interessensgesteuerte Variation der Analysetiefe" (1984, Projektleiter C.-R. Rollinger)."

(Rollinger85, p.27)

Die Ergebnisse dieser grundlagenorientierten Vorläuferprojekte bilden die Basis für die gegenwärtigen, z.T. entwicklungsorientierten Projekte. Diese neueren KIT-Projekte sind:

1. KIT-BACK:

"Das Ziel des Gesamtvorhabens ist die Entwicklung und prototypische Implementation eines komplexen Systems zur Verwaltung von grossen Daten- und Wissensbeständen mit eingeschränktem natürlichsprachlichem Zugang. Das an der TU-Berlin am 1.1.1985 begonnene Teilprojekt hat die Entwicklung und prototypische Implementation eines KBMS [**8] zum Ziel. Hierbei wird ein maschinell handhabbarer Formalismus zur Repräsentation von Wissen entwickelt, der für den begrifflichen Anteil des Wissens eine Weiterentwicklung von KL-ONE [**9], für den referentiellen Anteil eine Weiterentwicklung von dem an der TU-Berlin entwickelten Formalismus SRL [**10] und für den inferentiellen Anteil eigene, an partitionierten semantischen Netzen orientierten Formalismen beinhalten wird. Besonderer Schwerpunkt wird die Repräsentation unvollständigen und unsicheren Wissens sein. Durch dieses KBMS soll eine Basis geschaffen werden, die auch in den weiteren, in der KIT-Projektgruppe angesiedelten Projekten als Wissensverwaltungssystem und komplexe Inferenzmaschine zur Verfügung steht." (Rollinger85, p29)

Die Projektlaufzeit dauert von 1984-1989. Projektleiter ist Dr. K. v. Luck und Projektträger ist ESPRIT. (ESPRIT = Europäisches Strategisches Programm für Forschung und Entwicklung auf dem Gebiet der Informationstechnologie).

2. KIT-NATAN:

"Ziel des Projekts ist ein System, das auch dem gelegentlichen Benutzer den Zugang zu relationalen DB-Systemen ermöglicht. Der Übergang von der Anfrage in natürlicher Sprache zur DB-Schnittstellensprache erfolgt durch einen wissensbasierten Transformationsprozess. Dazu wird zum einen eine logisch-semantische Zwischenrepräsentation benötigt. In NATAN wird dafür die Sprache SRL verwendet und weiterentwickelt. Zum anderen muss für die automatische Anfragebehandlung regelhaftes Wissen über den Anwendungsbereich und die DB-Struktur bereitgestellt werden, das herkömmliche DB-Systeme nicht bzw. nicht explizit enthalten (Weltwissen und DB-Metawissen). Schliesslich werden Verfahren entwickelt, mit denen dieses regelhafte Wissen teilweise automatisch erzeugt werden kann. Für das Parsing wird ein Modell auf der Basis von "Wortartexperten" entwickelt. Mit Wortartexperten können die Vorteile des wortorientierten Links-Rechts-Parsens aus dem Word-Expert-Ansatz auf ein integriert syntaktisch-semantisches Verfahren übertragen werden, wobei das Wissen über Wortstämme und deren Verwendungsformen getrennt verwaltet werden kann." (Rollinger85, p.30)

Im Zusammenhang mit KIT-NATAN ist zu erwähnen, dass sich die als Antwort auf den Fragebogen zugesandten Arbeitsberichte von M. Eimermacher (Eimermacher83a,b,c) im Detail mit dem integriert syntaktisch-semantischen Verfahren befassen, das auf die im zitierten Projektbescheid erwähnten "Wortartexperten" zurückgreift. Das "Word-Expert-Parsing" wurde bereits im Vorläuferprojekt "Kognitive Verfahren zur Informations-Extraktion und Zusammenfassung aus Texten" benutzt. Beim synoptischen Vergleich der Antworten wird näher darauf eingegangen (TEIL 3).

Das andere Parsingverfahren, das im Projektrahmen KIT verwendet wird, nämlich der von Kilbury (Kilbury85) entwickelte GPSG-basierte Parser nach Earley/Shieber findet im Teilprojekt

3. KIT-NASEV

Anwendung. Dabei handelt es sich um eine Ergänzung zur EUROTRA Forschung. Maschinelle Übersetzung soll aus KI-Sicht behandelt werden. Die dazu benötigte formale Grammatik des Deutschen soll auf der Basis der Generalized Phrase Structure Grammar (GPSG) erstellt werden. Auf diesen Ansatz gehe ich nicht ein, weil das Thema der vorliegenden Arbeit nicht die MU ist.

4. KIT-Lerner

Auch auf dieses Projekt gehe ich nicht näher ein, weil es ebenfalls nicht in den Rahmen der vorliegenden Arbeit gehört.

Zum Thema Implementierung bemerkt (Rollinger85):

"Die Implementierungsarbeiten werden (wie seit 1981) in PROLOG erfolgen, in mehreren Umgebungen: IBM 4381-2, VAX 780 und (hoffentlich) Symbolics."
(Rollinger85, p.28)

Die Antworten.

Die Antwort wurde in Form eines Briefes und mit 4 ausführlichen Arbeitsberichten erteilt. Der Brief lautet:

"In der Projektgruppe KIT werden zur Zeit zwei Parserverfahren verwendet. J. Kilbury entwickelt einen GPSG-basierten Parser nach Earley/Shieber (s. Report 24); mein Ansatz verbindet die wortorientierte Verarbeitung des Word-Expert-Parsings mit eigenständigen Syntax-Experten (s. Report 15 und IAB 8). Für den GPSG-Ansatz gibt es bisher noch keine morphologisch-lexikalische Komponente; zum modifizierten Word-Expert-Parser muss ich auf IAB9 verweisen, da der Fragebogen sich zum grossen Teil auf Implementierungsdetails bezieht, von denen wir in PROLOG abstrahieren können. Z.B. interpretiert PROLOG eine Menge von "Fakten" auch als Datenbank, obwohl sie als gewöhnliche Datei oder Modul eingegeben werden kann. Ausserdem verzichten wir noch weitgehend auf platzsparende Codierungen im Lexikon. Die lexikalische Komponente ist inzwischen in einer Reihe von Punkten weiterentwickelt worden, die jedoch nichts am prinzipiellen Aufbau ändern. Z.B. sind die Wortstämme im Lexikon in Buchstabenlisten zerlegt, so dass eine passende Wortform beim Lexikonzugriff (als Nebenprodukt) in Stamm und Endung zerfällt. Ausserdem ist das Verfahren auf Affixe und Suffixe erweitert, dass beliebige Komposita verarbeitet werden können. Als Rechner benutzen wir eine IBM4381, Zahl der Stammformen zur Zeit: einige hundert."
(Eimermacher85)

Literaturhinweise.

1. M. Eimermacher, Ein Word Expert Parser in PROLOG - Integration von Syntax und Semantik. In: KIT Interner Arbeitsbericht 8 (IAB8), TU Berlin, Inst. f. angew. Informatik, Dez. 1983)
2. M. Eimermacher, Ein Word Expert Parser in PROLOG - Die lexikalische Komponente. In: KIT IAB9, TU Berlin, Dez 1983.
3. M. Eimermacher, Parsing in PROLOG - Eine semantische Komponente. In: KIT-Report 15, TU Berlin, Dez. 1983
4. J. Kilbury, Chart Parsing and the Earley Algorithm. In: KIT-Report 24, TU Berlin, Jan 1985.

2.2.3 Das Lexikon des Systems EVAR.

Dem allgemeinen Lexikon (genannt Null-Lexikon) wird über Präprozessoren die jeweils benötigte Informationsmenge entnommen und als Speziallexikon lokal für die entsprechende Systemkomponente aufbereitet.

Das Projekt.

EVAR, Erkennen-Verstehen-Antworten-Rückfragen, ist ein Projekt an der Universität Erlangen-Nürnberg am Lehrstuhl für Informatik 5 - Mustererkennung. Auf der GWAI-81 (German Workshop on Artificial Intelligence) wurde das Projekt durch H.-W. Hein zusammenfassend und perspektivisch vorgestellt:

"In 1979 we started the development of an experimental German speech understanding system [**11]. It should be speaker independent and accept connected speech of good signal quality. All components and their underlying models should be independent of each other as far as possible. This will enable an easy exchange of components and various experiments in constructing different systems out of existing modules.

Understanding strategy will be tailored to parallel computing purposes. In the first project phase parallelism is simulated on software level, in a second phase the system will be implemented on a parallel computer."
(Hein81,p277)

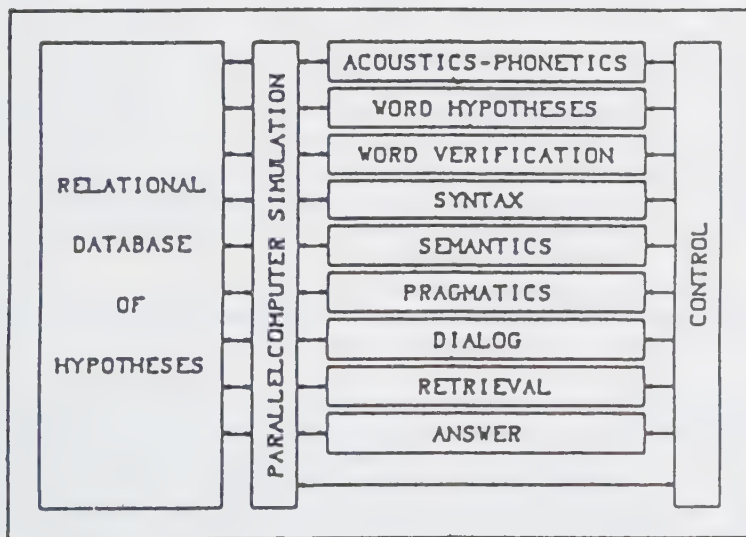
Der folgende Überblick basiert auf (Hein81) und (Niemann85). Das System EVAR ist wissensbasiert, das Lexikon ist eine von mehreren Wissensbasen. Das Wissen ist nach linguistischen Kriterien in verschiedene, autonome Wissensmodule unterteilt, jeder Modul läuft im Idealfall auf einem eigenen Prozessor eines Parallelcomputers, alle Module kommunizieren über eine relationale Hypothesenbank. Die Module waren ursprünglich folgende:

Akustik - Phonetik
 Phonologie - Prosodie
 Lexikon (Worterkennung)
 Syntax
 Semantik
 Pragmatik
 Dialog
 Strategie

Dieser Systemaufbau wurde jedoch in den Modulen Lexikon (=Worterkennung, nicht zu verwechseln mit dem allgemeinen Lexikon), sowie Retrieval und Answer erweitert:

"The system structure adopted for EVAR is according to a stratified linguistic model similar to the one given in (Winograd83), and the coordination of system modules is via a centralized data base and a control strategy as suggested first in HEARSAY II."
 (Niemann85, p273)

Folgende Abbildung zeigt die Systemarchitektur (Niemann85, p274):



"The two modules Word Hypothesatation and Word Verification in Fig. 2a [das entspricht der Abbildung der Systemarchitektur, H.H.] formerly were combined in only one module Lexicon. Furthermore, the module application [in der ursprünglichen Einteilung mit "Dialog" bezeichnet, H.H.] meanwhile was split into the three modules Dialog, Retrieval, and Answer."
 (Niemann85, p. 273)

Der Strategiemodul (Control) nimmt innerhalb dieses Systems von Experten (wie die Module auch genannt werden) einen besonderen Platz ein, weil er mit Steuerinformation den Ablauf der anderen Module beeinflusst.

Die Hypothesenbank ist in Analogie zu den linguistischen Abstraktionsebenen der Module ebenfalls modular aufgebaut. Die Hypothesen werden u.a. auf der phonologischen Ebene (Phon-Hypothesen), auf der Wortebene (Wort-Hypothesen), auf der syntaktischen Ebene (Satzstruktur-Hypothesen) und auf der semantischen Ebene (Satzbedeutungs-Hypothesen) gebildet.

Jeder Wissensmodul ist in der Lage, mit Hilfe seines Spezialwissens

- a) eine Hypothese zu bilden,
- b) eine Hypothese zu bewerten.

Gerade wegen des modularen Aufbaus sowohl der Hypothesenbank als auch der Wissensexperten ist es denkbar, dass weitere Wissens Ebenen repräsentiert werden können, ohne dass das ganze System verändert werden muss.

Grundsätzlich schreitet die Analyse von tieferen Abstraktionsebenen zu höheren, aber diese bottom-up Strategie kann effizienter gestaltet werden, wenn im Analyseprozess frühzeitig high-level Hypothesen gebildet werden können, welche unnötige low-level Alternativen vermeiden helfen. Als unterste Abstraktionsebene gelten die Daten über die eingegangenen akustischen Signale.

Der Strategiemodul optimiert den Systemzustand zwischen den kritischen Zuständen der Generation und Berücksichtigung von zu vielen bzw. zu wenigen Hypothesen. Zu viele Hypothesen bewirken eine Überlastung des Systems, zu wenige führen zu falschen Resultaten. Der Strategiemodul verfügt deshalb über Systemwissen, Optimierungsalgorithmen, Schnittstellen zu den linguistischen Expert-modulen und -wie alle anderen Module auch- über eine Schnittstelle zur Hypothesenbank. Mit Steuerparametern kann der Strategiemodul eingestellt werden.

Innerhalb der linguistischen Wissensbasen nimmt das Lexikon insofern eine Sonderstellung ein, als im Lexikon Wissen zu allen Gebieten der übrigen linguistischen Module codiert ist.

Diese Sonderstellung schlägt sich in der Konzeption des sog. Null-Lexikons nieder. Das Null-Lexikon ist ein Bestand von wortbezogenen linguistischen Daten. Dabei wird die Tatsache berücksichtigt, dass die "lexikalische Information" als solche gar nicht existiert, sondern immer eine nach linguistischen Kriterien differenzierbare Sammlung verschiedener wortbezogener Informationen ist. Diese Informationen teilen sich in die graphematische, phonetische Repräsentation der Wortform (In EVAR wird ein Vollformen-Lexikon verwendet) und in die der Wortform zugeordneten Informationen zu morphologischen, syntaktischen, und semantischen Kategorien.

Das Null-Lexikon wird von den übrigen linguistischen Modulen nicht direkt benutzt, weil vom Standpunkt jedes Spezialmoduls im Null-Lexikon zu viele für den Spezialisten nicht relevante Informationen vorhanden sind. Die Verbindung Nulllexikon - Expertmodul geschieht deshalb über einen Präprozessor, dessen Aufgabe darin besteht, für jeden Spezialmodul nur die benötigte Information aus dem Null-Lexikon zur Verfügung zu stellen, in einer

Repräsentationsform, die der Form im Spezialmodul entspricht. Diese Lexikonarchitektur garantiert einerseits die Erhaltung der konzeptionellen Informationseinheit "Wörterbuch" und andererseits die modulare Wissensstruktur der gesamten Systemarchitektur. Das Wörterbuch kann deshalb nach lexikographischen Kriterien konsistent gepflegt werden, und die Spezialmodule werden nicht mit irrelevanter Information belastet.

Als Beispiel der Lexikonbenutzung kann die Arbeitsweise des Moduls Syntax betrachtet werden. In (Niemann85) wird der Syntaxmodul in Stichworten wie folgt beschrieben:

```
input: verified word hypotheses

determine triples of words with syntactically allowed
word classes

determine larger sequences of words with syntactically
allowed word class sequences

select some high scoring sequences of words

perform syntactic analysis with score-dependent rules

generate top-down word hypotheses in gaps of syntactic
analysis

output: syntactic constituents of sentences represented
in a syntax analysis tree
```

Figure 9. 'Syntax-Modul (EVAR)': Modulaufbau (Niemann85, p.289)

Diese Beschreibung entspricht einer neueren, bezüglich der ursprünglichen Konzeption des Syntaxmoduls revidierten Beschreibung der Arbeitsweise dieses Moduls.

"The early version of the Syntax module is presented in (Hein82), (Niemann82). It used a fairly elaborate ATN-grammar of German language and an island parser. The problem with this approach is the combinatoric explosion due to the fairly large number of word alternatives and the rigid constraints imposed by the ATN-rules, for example the case-number-gender agreement in noun groups. Therefore, we plan a redesign of the module according to Fig. 7 [entspricht obiger Modulbeschreibung in Stichworten, H.H.]. The two main points are that a fast prescanning of words will be done to detect a few sequences of words belonging to syntactically allowed word classes and that the rigidity of syntactic constraints will be dependent on the score of words. For example the case-number-gender agreement may depend only on some endings of words which are hard to recognise and, therefore, this should be checked only for very

high scoring words."
(Niemann85, p.288)

Der syntaktische Parser des Syntaxmoduls benutzt beim Zugriff auf lexikalisches Wissen nicht das Nullexikon sondern das durch einen Präprozessor aufbereitete Speziallexikon für den Syntaxexperten. Auch für die anderen linguistischen Expert-Module wird über den Präprozessor genau diejenige Nullexikon-Information extrahiert, die dem Modul dient. Für den Syntaxexperten sind das morphosyntaktische lexikalische Informationen und für den Semantikexperten lexikalische Valenz- und Tiefenkasusinformationen.

Die Antworten.

1. FRAGE:

Verwenden Sie mehrere Typen von Lexika? (Vollformenwörterbuch/ Grundformenwörterbuch/ Stammwörterbuch/ andere)

ANTWORT:

Vollformenwörterbuch

2. FRAGE:

Welchen Zwecken dient das Lexikon? (Lemmatisierung/ Syntax/ Semantik/ maschinelle Übersetzung/ Information, Dokumentation/ Künstliche Intelligenz/ andere)

ANTWORT:

Syntax, Semantik, Künstliche Intelligenz
andere: automatische Sprachverarbeitung (gesprochene Sprache)

3. FRAGE:

Welches Wortklassenschema wird zugrundegelegt?

- Wortarten (Tabelle, ev. Abkürzungen).
- Welche Wortarten unterteilen Sie in welche Unterklassen?
- Welche Wortarten fassen Sie zu welchen Oberklassen zusammen?

ANTWORT:

Vgl. Veröffentlichung in: Sprache und Datenverarbeitung, 1-2/1984
p. 54-63. (Bibliographiert als Brietzmann84b)

4. FRAGE:

Welche Informationsfelder werden pro Eintrag codiert? (zusammengehörende Informationsfelder (Codes) als solche markieren) Falls die Einträge nach Wortarten differenziert codiert sind, dies in analogen Tabellen darstellen.

ANTWORT:

Aus- spra- che	Nor- mal- form	Stamm, Flex.- art	morpho synt. Merk- male	Wort- art	Se- mant. Klassi- fikat.	Kasus- rahmen
----------------------	----------------------	-------------------------	----------------------------------	--------------	-----------------------------------	------------------

5. FRAGE:

Welchen Rechner benutzen Sie?

ANTWORT:

VAX 11-730, Cyber 895

6. FRAGE:

Ist Ihr Verfahren fileorientiert oder datenbankorientiert?

ANTWORT:

momentan fileorientiert

7. FRAGE:

Aus welchen Feldern ist ein logischer Record aufgebaut?

ANTWORT:

Textfile

Die nächsten Teilfragen entfallen.

Die 8. Frage entfällt.

9. FRAGE:

Sind ausser der codierten Lexikoninformation noch andere Informationen codiert? (Affixlisten/ andere)

ANTWORT:

--

10. FRAGE:

In welcher Programmiersprache operieren Sie über dem Wörterbuch?

ANTWORT:

Zukünftig in Lisp

11. FRAGE:

Mit welchen Zugriffsalgorithmen operieren Sie? (VSAM/ hash/ B-Baum/ andere)

ANTWORT:

noch unklar

12. FRAGE:

Welche Felder können als Zugriffsschlüssel verwendet werden? (nur ein Schlüssel/ alle Felder/ andere)

ANTWORT:

derzeit nur ein Schlüssel. Wortnummer

13. FRAGE:

Wie gross ist Ihr Wörterbuch? (Anzahl Einträge/ Anzahl Grundformen/ Anzahl abgeleitete Formen/ andere Grössemerkmale)

ANTWORT:

ca 5000 Einträge

14. FRAGE:

Maximale Wortlänge?

ANTWORT:

beliebig

15. FRAGE: Maximale Anzahl von Informationskategorien (Codes) pro Lexem?

ANTWORT:

beliebig

16. FRAGE:

Besteht die Möglichkeit, Einträge zu verknüpfen? (Verknüpfung Lemma- Formen/ Thesaurusrelationen/ Frames/ andere

ANTWORT:

Verknüpfung Lemma-Formen

17. FRAGE:

Welche LDV-Programme (oder Systeme) benutzen Ihre Wörterbücher?

ANTWORT:

--

18. FRAGE:

existiert eine Schnittstelle zur DIN 2341 (z.B. MATER)? [**7]

ANTWORT:

--

19. FRAGE:

Welche konventionellen Wörterbücher dienen als Grundlage? (für den Wortschatz/ für die Codierung/ andere Informationen)

ANTWORT:

- Grundwortschatz Deutsch
- Valenzwörterbücher
 - a) Helbig/Schenkel
 - b) Schumacher (Institut f. dt. Sprache)
- Bedeutungswörterbuch (DUDEN)

20. FRAGE:

Ist es Ihnen möglich, einen typischen Auszug aus dem Wörterbuch und eine Erklärung des Codes beizulegen?

ANTWORT:

--

21. FRAGE:

Literaturhinweise:

ANTWORT:

--

Hinweise.

Diese Hinweise wurden zusätzlich zu den Antworten auf den Fragebogen notiert.

1. Projekt "Kumulative Wortdatenbank" am IKP, Uni Bonn (Jan Brustkern).

2. Studienarbeiten zum Thema Lexikon.

Vergl. dazu Kapitel 3.3.2

*

Die in der Antwort zu Frage Nr.3 erwähnte Veröffentlichung von A. Brietzmann enthält folgendes Wortklassenschema:

ADJ	(attributives) flektiertes Adjektiv
ADJU	unflektiertes Adjektiv
ADV	Adverb
AUXH	Hilfsverb HABEN
AUXS	Hilfsverb SEIN
AUXW	Hilfsverb WERDEN
DET	Determinans (possessivische, indefinite, demonstrative, definite Artikelwörter)
EPAR	emotionale Partikel („JA, ETWA, HALT“)
GRADPAR	Gradpartikel („SEHR, SO, ÄUSSERST“)
INFSUBJ	Infinitivsubjunktion („UM, OHNE, ANSTATT“)
KONJ	Konjunktion
KOORD	Koordination
MADV	Modaladverb („ANSCHEINEND, GEWISS“)
MVERB	Modalverb
N	Nomen
NEG	Negationspartikel („NICHT, KEINESFALLS“)
NPR	Eigennamen (nomen proprium)

Figure 10. 'Wortklassen (EVAR)': (Brietzmann84b, p62)

ORDZAHL	Ordinalzahl
PRAEP	Präposition
PRAEPRAEP	'Prä'präposition („BIS zu, BIS auf“)
PRON	Pronomen (satzgliedfähige Personal-, Indefinit-, Demon- strativpronomina)
REFLPRON	Reflexivpronomen
RELPRON	Relativpronomen
SUBJ	Subjunktion
VERB	Vollverb
VERBPRAE	(abspaltbares) Verbpräfix
VERPAR	Vergleichspartikel („ALS, WIE“)
WADV	Frageadverb („WANN, WO“)
WDET	Fragedeterminans („WEL- CHER“)
WPRON	Fragepronomen („WER, WAS“)
ZAHLWORT	Kardinalzahl

Forts. 10. 'Wortklassen (EVAR)': (Brietzmann84b, p. 22)

Die Lexikoneinträge in den verschiedenen lokalen Lexika des Systems EVAR unterscheiden sich von der Eintragsstruktur im Nullexikon. Hier ein Beispiel:

"Der Semantik-Modul bezieht sein lexikalisches Wissen aus einem eigenen Wörterbuch, das nur die Semantik-Einträge der Grundformen enthält. Ein Lexikon-Eintrag hat prinzipiell folgenden Aufbau: (Brietzmann84a, p. 93)

```

wortart_1:

    bedeutung_1:
        merkmalf_1
        :
        :
        merkmalf_i
        :
        :
    bedeutung_m:
        merkmalf_1
        :
        :
        merkmalf_j
        :
        :
wortart_n:

    bedeutung_1:
        merkmalf_1
        :
        :
        merkmalf_k
        :
        :
    bedeutung_k:
        merkmalf_1
        :
        :
        merkmalf_l

```

Figure 11. 'Semantik-Eintrag (EVAR)'

"Das lokale Wörterbuch der Semantik-Komponente wird durch einen Präprozessor aus dem Vollformen-Lexikon des Spracherkennungs-Systems erstellt".

"Da semantische Information nur bei Grundformen eingetragen ist, reduziert sich der Umfang des Lexikons gegenüber dem Gesamt-Wörterbuch sehr stark".
 (Brietzmann84a, p. 94)

2.2.4 Das Lexikon SADAW.

Das Lexikon SADAW stellt mit über 130'000 Einträgen, in denen hauptsächlich morphosyntaktische Kategorien codiert sind, ein lexikalisches Formenwissen dar, das in unterschiedlichen Bereichen wie "Maschinelle Übersetzung" und "Juristische Dokumentation" genutzt wird.

Das Projekt.

SADAW (Saarbrücker Deutsches Analyse Wörterbuch) wurde im Rahmen des Sonderforschungsbereichs 100 "Elektronische Sprachforschung" entwickelt. In diesem Rahmen erfuhren verschiedene Teilprojekte eine Förderung durch die Deutsche Forschungsgemeinschaft. Dazu gehörten u.a.:

- Elektronische Syntaxanalyse
- Automatische Übersetzung
- Automatische Lemmatisierung

Vom BMFT wurde über die GID gefördert:

-Juristische Dokumentenanalyse (seit 1980 auch in Saarbrücken) .

In Saarbrücken befindet sich auch der grösste Teil der deutschen EUROTRA Forschung.

In allen diesen Projekten wird (bzw. wurde) das Lexikon SADAW (bzw. sein Vorläufer SDW) eingesetzt. Das Lexikon SADAW ist so konzipiert, dass es in thematisch unterschiedlichen Projekten Verwendung finden kann. Das gilt aber nur für Projekte mit bestimmten Voraussetzungen, denn von der Implementation her betrachtet, ist das Lexikon SADAW nicht so flexibel:

"Um das Wörterbuch lesen, erweitern, korrigieren, etc. zu können"

benötigt man

"ein umfangreiches Softwarepaket (=UPDATEN), das in FORTRAN geschrieben ist und nur auf einem grossen Siemensrechner zur Verfügung steht. An Speicherplatz benötigt SADAW 10 MB, UPDATEN 720 KB. Um Updaten laufen lassen zu können, benötigt man 320 KB."
(H.-D. Luckhardt) [**13]

Detaillierte Beschreibungen über Aufbau und Inhalt des Lexikons SADAW findet man u.a. in folgenden Publikationen:

- Eggers/Luckhardt/Maas/Weissgerber (hg), Salem, Ein Verfahren zur automatischen Lemmatisierung deutscher Texte. (Eggers80)
- Linguistische Arbeiten des SFB100, Heft 3.2, 1980
- (Hess83), p.83-98 sowie p.190-194
- Auf Fragen der Integration der Komponente SADAW im System Judo geht H. Werner ein: "Eine Sprachdatenbank: Zur Gestaltung der Wörterbucharbeit" (Werner80). Vergl. dazu auch: (Zimmermann78).
- Über neuere Anwendungen im Zusammenhang mit MÜ siehe: H.J. Weber, Ein dreistufiges Lexikonsystem zur maschinellen Analyse von semantischen Textstrukturen. In: Textstrukturen und maschinelle Übersetzung, Arbeitsberichte des Projekts A3 des sfb100, EMA Nr. 6.2, Saarbrücken 1984.

SADAW ist nicht das einzige Lexikon, welches im sfb100 verwendet wird.

Die Antworten.

1. FRAGE:

Verwenden Sie mehrere Typen von Lexika? (Vollformenwörterbuch/ Grundformenwörterbuch/ Stammwörterbuch/ andere) ANTWORT:

Grundformenwörterbuch

2. FRAGE:

Welchen Zwecken dient das Lexikon? (Lemmatisierung/ Syntax/ Semantik/ maschinelle Übersetzung/ Information, Dokumentation/ Künstliche Intelligenz/ andere)

ANTWORT:

(Lemmatisierung), Syntax, Semantik, maschinelle Übersetzung, Information, Dokumentation (System CTX (Prof. Zimmermann, Saarbrücken) basiert auf SUSY.).

3. FRAGE:

Welches Wortklassenschema wird zugrundegelegt?

- Wortarten (Tabelle, ev. Abkürzungen).

Teilantwort:

--

- Welche Wortarten unterteilen Sie in welche Unterklassen?

Teilantwort:

- Welche Wortarten fassen Sie zu welchen Oberklassen zusammen?

Teilantwort:

- Verben
- Adjektive
- Substantive.
- Funktionswörter

4. FRAGE:

Welche Informationsfelder werden pro Eintrag codiert?
(zusammengehörende Informationsfelder (Codes) als solche markieren)
Falls die Einträge nach Wortarten differenziert codiert sind, dies
in analogen Tabellen darstellen.

ANTWORT:

--

5. FRAGE:

Welchen Rechner benutzen Sie?

ANTWORT:

Siemens 7760

6. FRAGE:

Ist Ihr Verfahren fileorientiert oder datenbankorientiert?

ANTWORT:

fileorientiert

7. FRAGE:

- Aus welchen Feldern ist ein logischer Record aufgebaut?

Teilantwort:

--

- Welche Recordarten benutzen Sie

Teilantwort:

--

- Entsprechen Ihre Recordarten den Wortarten?

Teilantwort:

ja

Die 8. Frage entfällt.

9. FRAGE:

Sind ausser der codierten Lexikoninformation noch andere
Informationen codiert? (Affixlisten/ andere)

ANTWORT:

Affixlisten

Präfixe, Suffixe, Fugen sind ebenfalls Einträge.

10. FRAGE:

In welcher Programmiersprache operieren Sie über dem Wörterbuch?

ANTWORT:

FORTRAN

Der Computerlinguist benutzt nur bestimmte Unterprogramme.

Wie's drinnen aussieht, braucht nur der Softwaremann zu wissen.

11. FRAGE:

Mit welchen Zugriffsalgorithmen operieren Sie? (VSAM/ hash/ B-Baum/ andere)

ANTWORT:

B-Baum

12. FRAGE:

Welche Felder können als Zugriffsschlüssel verwendet werden? (nur ein Schlüssel/ alle Felder/ andere)

ANTWORT:

nur ein Schlüssel

13. FRAGE:

Wie gross ist Ihr Wörterbuch? (Anzahl Einträge/ Anzahl Grundformen/ Anzahl abgeleitete Formen/ andere Grössenmerkmale)

ANTWORT:

--

14. FRAGE:

Maximale Wortlänge?

ANTWORT:

36 Buchstaben

15. FRAGE: Maximale Anzahl von Informationskategorien (Codes) pro Lexem?

ANTWORT:

256 Byte + 2 Wortlaute

16. FRAGE:

Besteht die Möglichkeit, Einträge zu verknüpfen? (Verknüpfung Lemma- Formen/ Thesaurusrelationen/ Frames/ andere

ANTWORT:

nein

17. FRAGE:

Welche LDV-Programme (oder Systeme) benutzen Ihre Wörterbücher?

ANTWORT:

Unser eigenes, innerhalb von SUSY

18. FRAGE:

existiert eine Schnittstelle zur DIN 2341 (z.B. MATER)? [**7]

ANTWORT:

--

19. FRAGE:

Welche konventionellen Wörterbücher dienen als Grundlage? (für den Wortschatz/ für die Codierung/ andere Informationen)

ANTWORT:

für den Wortschatz:
deutsch: Wahrig
sonst Häufigkeits-WB, 'Grundwortschatz',
Wörter aus unseren Korpora

20. FRAGE:

Ist es Ihnen möglich, einen typischen Auszug aus dem Wörterbuch und eine Erklärung des Codes beizulegen?

ANTWORT:

--

21. FRAGE:

Literaturhinweise:

ANTWORT:

Linguistische Arbeiten des SFB, Heft 3.2, 1980.

Hinweis.

Dieser Hinweis wurde zusätzlich zu den Antworten auf den Fragebogen notiert:

Prof. Lenders, Uni Bonn. Projekt über deutsche masch. Lexika.

Abgesehen von den in der Teilantwort zu Frage Nr.3 erwähnten Oberklassen von Wortarten (Verben, Adjektive, Substantive, Funktionswörter) sind die Wortarten in SADAW nach folgenden Analysewortklassen eingeteilt: (Ein ähnlicher Überblick über diese Analysewortklassen findet sich auch in (Eggers80,p34ff). Die Klassen werden mit den Benennungen nach der DIN MATER [**14] verglichen.

SADAW		MATER	
Kat.	Benennung/Inhalt	Kat.	Benennung/Inhalt
SUB	Substantiv	21o	Substantiv
ADJ	Attributiv	221:	Numerale:
	a) attributive Partizipien	D	Ordnungszahl
	b) Ordinalzahlen		
	c) Formen von ein, beid		
MASO	manch, solch	-	
ANO	Adverb adnominal, Partikel	242:	Partikel:
	a) Verstärkung	B	verstärkend
	b) Restriktionspartikel	A	begrenzend
PER	Personalpronomen	222:	Pronomen:
	a) alle üblichen Personalpronomen	A	Personalpronomen
	b) bei "wir danken dem, der"		
	c) wer, was, wessen - interrogativ	F	Interrogativ-Pronomen
	d) relativisch	E	Relativpronomen
	e) reflexiv	B	Reflexivpronomen
POS	Possessivpronomen	222:	Pronomen:
		C	Possessivpronomen
DEM	Demonstrativpronomen	222:	Pronomen:
		D	Demonstrativpronomen
IND	Indefinitpronomen	222:	Pronomen:
		F	Indefinitpronomen
VIEL	viel und wenig	24o:	Adverb:
		A	Qualitativ
ALL	alle Formen von 'all', aber nicht 'all' selbst	-	
BEID	alle Formen von beid	-	
ARTB	bestimmter Artikel	211:	Artikel:
		A;B	best. mask.sing; best. fem. sing;
		C;D	best. neutr. sing; best. plural
ARTU	unbestimmter Artikel	211:	Artikel:
		E;F	unbest. mask.; unbest. fem.
		G	unbest. neutr.
PRP	alle üblichen Präpositionen	212	Präposition
POP	Postposition	212:	Präposition:
		J	nachgestellt
SELB	postponierte Partikel wie 'selbst'	-	
NUM	Zahlwörter	221:	Numerale:
	a) Kardinalzahl	A	Kardinalzahl
	b) Ziffern		
ADV	alle Adverbien, prädikative	24o	Adverb
	Adjektive, sonstige	22o:	Adjektiv
	Partikel	W	prädikativ

Figure 12. 'Analysewortklassen (SADAW)': (Hess83,p191)

SADAW		MATER	
Kat.	Benennung/Inhalt	Kat.	Benennung/Inhalt
		242: C;D	Partikel: gegenüberstellend; grammatikali- siert
SAV	Satzadverb	-	
FIV	finites Verb=Verb- formen im Präsens, Prä- teritum, Konjunktiv, Indikativ, Imperativ	230: AJ AK AF AE AI	Verb: Präsens Imperfekt Konjunktiv Indikativ Imperativ
INF	Infinitiv	230: AA	Verb: Infinitiv
IZU	Infinitiv mit einge- schlossenem 'zu'	-	
PTZ	Partizip II	230: AD	Verb: Partizip II
VZS	Verbzusatz=abtrennbare Verbpräfixe	230: AX	Verb: Präfix abtrennbar
ZU	Wortklasse 'zu' in: "er wagte es nicht zu fra- gen"	-	
ITJ	Interjektion	244: A;B C	Interjektion: Empfindung; Anruf/Aufforderung; Lautnachahmung
FRE	Fremdwort	-	
UKO	unterordnende Kon- junktion		
DASS	Konjunktion 'daß'	243: D	Konjunktion: subordinierend
OB	Konjunktion 'ob'		
UOA	Konjunktion vom Typ 'um, ohne, anstatt, statt'		
SEN	Satzende	-	
NKO	Konjunktion, nebenord- nend	243: A	Konjunktion: kopulativ

Figure 13. 'Forts. Analysewortklassen (SADAW)': (Hess83, p192)

2.2.5 Das Lexikonsystem ALEXSYS.

Dieses Lexikonsystem ist eine kumulative Wortdatenbank, d.h. es besteht

- a) aus mehreren Einzellexika, die in der Originalform rekonstruierbar sind, und
- b) aus dem kumulierten Lexikon, welches eine Eintragsstruktur aufweist, die mit den Einzellexika kompatibel ist.

Das Projekt.

Das Akronym ALEXSYS bedeutet ALLgemeines LEXikonSYStem und bezeichnet ein Projekt, das seit dem 1.3. 1982 vom BMFT (Bundesminister für Forschung und Technologie) im Rahmen der Förderung von Information und Dokumentation durch den Projektträger GID (Gesellschaft für Information und Dokumentation) gefördert wurde.

"Die Zielsetzung, aus insgesamt 12 maschinenlesbar verfügbaren Einzellexika des Deutschen ein kumuliertes Lexikon zu erstellen und dieses in eine Wortdatenbank zu überführen" (Heinze85, p.38)

wurde am 28.11.1984 mit der Vorführung der SESAM Wortdatenbank des Deutschen im Institut für deutsche Sprache (Mannheim) erreicht.

Die kurze Projektbeschreibung basiert auf:

- K. Hess, J. Brustkern, W. Lenders, Maschinenlesbare deutsche Wörterbücher. Dokumentation, Vergleich, Integration. Tübingen 1983 (Hess83).

- J. Brustkern, W. Schulze, Aufbau einer Wortdatenbank für das Deutsche, Sonderdruck SDV, 1983 (Brustkern83)

- W. Schulze, G. Heinze, Die morphosyntaktische Komponente in der Wortdatenbank des Deutschen. (Schulze83)

- G. Heinze, J. Brustkern, Das Lexikonsystem ALEXSYS. (Heinze85)

Das Lexikonsystem ALEXSYS am IKP (Institut für Kommunikation und Phonetik der Universität Bonn) und die Wortdatenbank SESAM am IDS haben den gleichen Datenbestand. Die folgende Tabelle aus (Heinze85) zeigt Umfang und Herkunft der im System enthaltenen Lexika:

Name	Umfang	Lexemtyp	Herkunft
LIMAS	129.510	diverse Formen	Forschungsprojekt "Linguistische und Maschinelle Sprachverarbeitung"; 1965-1976, IKP, Universität Bonn
DTPREP	240	diverse Formen Präpositionen	LIMAS-Projekt
NOMPREP	1.115	diverse Formen Nomen-Präpositionen- Ebindungen	LIMAS-Projekt
VRBPREP	1.798	diverse Formen Verb-Präpositionen- Ebindungen	LIMAS-Projekt
MACKENSEN	113.578	Grundformen	Auszug des "Deutschen Wörterbuches" von L. Mackensen, 1955; in einer Bearbeitung von G. Kandler und K. Bünting
MOLEX	24.087	diverse Formen	Forschungsprojekt PLIDIS ("Problemlösendes Informationssystem mit Deutsch als Interaktionsprache"), 1972-1976, IDS, Mannheim
SDW	94.036	diverse Formen	Saarbrücker Deutsches Wörterbuch, entwickelt für die automatische Lemmatisierung, 1970-1975, SFB 100, Universität Saarbrücken
SADAW	113.135	diverse Formen	Saarbrücker Deutsches Analyse-Wörterbuch entwickelt als Nachfolge des SDW, seit 1976, SFB 100, Saarbrücken
CONDOR	843	Grundformen Funktions- wörter	Funktionswörterbuch des Dokumentationssystems CONDOR, entwickelt bei der SIEMENS AG München (1972-1981)
VALEX	ca. 1600 Einträge zu 425 Verben	diverse Formen Verben	Kleines Valenzlexikon Deutscher Verben, U.Engel, H. Schumacher, IDS Mannheim 1976
WAHRIG (Verben)	4.280	Grundformen Verben	Projekt zur Beschreibung der Bedeutungsstrukturen der deutschen Sprache von G. Wahrig, Mainz, 1974-1978

Figure 14. 'Datenbasis des kumulierten Lexikons': ALEXSYS und SESAM

Die in der Tabelle aufgeführten Datenbestände gelten sowohl für ALEXSYS als auch für SESAM. Der Unterschied zwischen den beiden Systemen ALEXSYS und SESAM liegt lediglich in der Implementation, wobei ALEXSYS ein in PL1 realisiertes File-System (ISAM) ist und die Wortdatenbank des Deutschen mit Hilfe von SESAM, einem Datenbankmanagementsystem der Firma Siemens, implementiert wurde.

Es handelt sich bei diesen beiden Wortdatenbanken nicht nur um eine Addition von Einzellexika, sondern um die Transformation der Eintragsstrukturen jedes Lexikons in ein gemeinsames Klassifikationssystem mittels Transformationslisten und einem aufwendigen Kumulationsverfahren.

Dabei wurde darauf geachtet, dass einerseits die Eintragskategorien der Einzellexika erhalten, bzw. rekonstruierbar blieben, und dass andererseits alle Eintragskategorien aller Einzellexika in ein übergeordnetes Eintragsschema passen. Für den Benutzer stellt sich dabei der Effekt ein, dass Daten sowohl in den Kategorien der Einzellexika als auch in den Kategorien der kumulierten Wortdatenbank genutzt werden können. Im Normalfall wird der Benutzer nur das kumulierte Klassifikationssystem kennen müssen, um nach einheitlichen Kriterien auf den gesamten Datenbestand zugreifen zu können.

*

Allerdings stellt das kumulierte Lexikon-System nach Ansicht der am Projekt Beteiligten nur ein Zwischenziel auf dem Weg zu einer vereinheitlichenden Wortdatenbank des Deutschen dar:

"Das integrierte Lexikon wird im Hinblick auf ein allgemeines vereinheitlichendes Lexikon als Fernziel weiter angestrebt; der kumulierte Aufbau kann als Vorstufe zu diesem Fernziel angesehen werden."

(Hess83, p. 217)

Erreicht ist dieses Ziel nicht, weil die verwendeten Einzellexika auch in ihrer Summe nicht zu allen Kategorien des gemeinsamen Wortklassensystems Informationen aufweisen und weil das gemeinsame System vom Projekt her direkt aus den Strukturen der Einzellexika abgeleitet werden musste. Integration würde deshalb bedeuten:

1. Beseitigung der Unvollständigkeiten hinsichtlich der Menge der Einträge und auch hinsichtlich der Vollständigkeit der einzelnen Einträge.

2. Erweiterung und z.T. Änderung der Integrationsstruktur.

"Besondere Unvollständigkeiten sind im Bereich der 'syntaktisch-semantischen Angaben' zu vermerken, da sich nur in 4 der insgesamt 12 Einzellexika Angaben zu dieser Informationsklasse befinden."

(Heinze85, p. 85)

*

Auf das Wortklassensystem und den Aufbau der Einträge wird im Anschluss an die Präsentation der Fragebogenantworten eingegangen.

*

Vom Standpunkt der Systemarchitektur betrachtet, ist das System ALEXSYS wie folgt aufgebaut:

Das Software-Paket

"besteht aus drei voneinander unabhängigen Komponenten:

1. Der Kumulationskomponente: Sie dient der schrittweisen Kumulation der Wortformen und der zugehörigen Angaben zu den Informationsklassen Wortklasse, Morphosyntax, Syntax und Semantik in den 12 Einzelllexika.
2. Der UPDATE-Komponente: Sie dient der interaktiven Pflege des kumulierten Lexikons.
3. Der RETRIEVAL-Komponente: Sie dient der interaktiven Wiedergewinnung von Datensätzen aus der jeweils aktuellen Fassung des kumulierten Lexikons.

Bei der Realisierung der RETRIEVAL- und der UPDATE-Komponente findet die Menue-Technik Verwendung".
(Heinze85, p38/39)

"Das Kernstück der Kumulationskomponente ist ein Paket von Batch-Programmen zur schrittweisen Kumulation der Wortformen und der zugehörigen informationsklassen-spezifischen Angaben in den 12 Einzelllexika."
(p.40)

Langfristig betrachtet ist das Konzept einer integrierten Wortdatenbank besonders interessant, wenn man den Zusammenhang zu experimentellen linguistischen Lexikonsystemen für Textanalyse und Wörterbuchherstellung beachtet. Solche Systeme sind etwa BONNLEX1 (Willée85) oder Teile der LDVLIB (Drewke75) in Zürich. Diese Systeme sind multifunktional und dienen u.a. der Wörterbuchherstellung, der wörterbuchunterstützten Textanalyse, der Lemmatisierung, Inhaltsanalyse und Konkordanzherstellung. Wird nun ein solches System mit einer Wortdatenbank gekoppelt, erweitern sich die Anwendungsmöglichkeiten beträchtlich.

Eine weitere Perspektive im Zusammenhang mit der Wortdatenbank ALEXSYS ergibt sich durch die Kombination der Datenbank mit PC-Software, etwa mit Textretrieval-Systemen und lexikographischen "Workstations". Ein Beispiel dafür ist das System BYU, welches am IKP für deutsche Benutzerverhältnisse adaptiert wurde (entwickelt wurde dieses Konkordanzsystem für PC an der Brigham Young University, Utah, 1985).

Allerdings ist hier einschränkend festzuhalten, dass dieser Trend zur Integration von PC-Software und Grossystemen nicht nur auf dieses Beispiel der Wortdatenbank beschränkt ist, sondern einer allgemeinen Tendenz in der Informatik entspricht.

Die Antworten.

1. FRAGE:

Verwenden Sie mehrere Typen von Lexika? (Vollformenwörterbuch/ Grundformenwörterbuch/ Stammwörterbuch/ andere)

ANTWORT:

Vollformen- und
Grundformenwörterbuch gemischt

2. FRAGE:

Welchen Zwecken dient das Lexikon? (Lemmatisierung/ Syntax/
Semantik/ maschinelle Übersetzung/ Information, Dokumentation/
Künstliche Intelligenz/ andere)

ANTWORT:

Generelle lexikalische Datenbasis, die für alle genannten
Bereiche verwendbar ist.

3. FRAGE:

Welches Wortklassenschema wird zugrundegelegt?

- Wortarten (Tabelle, ev. Abkürzungen).
- Welche Wortarten unterteilen Sie in welche Unterklassen?
- Welche Wortarten fassen Sie zu welchen Oberklassen zusammen?

ANTWORT:

das Wortklassenschema ist dem beigefügten Sonderdruck
SDV 2/1980 zu entnehmen s. 172 ff

4. FRAGE:

Welche Informationsfelder werden pro Eintrag codiert?
(zusammengehörende Informationsfelder (Codes) als solche markieren)
Falls die Einträge nach Wortarten differenziert codiert sind, dies
in analogen Tabellen darstellen.

ANTWORT:

siehe Hinweise nach dem Fragebogen.

5. FRAGE:

Welchen Rechner benutzen Sie?

ANTWORT:

IBM 370/168

6. FRAGE:

Ist Ihr Verfahren fileorientiert oder datenbankorientiert?

ANTWORT:

fileorientiert.

Wird als Datenbank am Institut für Deutsche Sprache, Mannheim,
aufgebaut

7. FRAGE:

- Aus welchen Feldern ist ein logischer Record aufgebaut?

Teilantwort:

Siehe Hinweise nach dem Fragebogen.

- Welche Recordarten benutzen Sie

Teilantwort:

4 ISAM-Files für die Wortartengruppen

- Nomen
- Verben
- Qualifier (Adj. und Adverbien)
- Funktionswörter

- Entsprechen Ihre Recordarten den Wortarten?

Teilantwort:

nein

8.FRAGE:

Falls datenbankorientiert:

- welchen DB-typ verwenden Sie? (CODASYL/ Baumhierarchie/ Inverted File/ Relational/ andere)

ANTWORT:

--

Teilfrage:

- Wie lautet die Bezeichnung Ihrer Datenbank als kommerziell erwerbbares DBMS?

Teilantwort:

beim IDS handelt es sich um eine SESAM-Datenbank der Firma Siemens.

Teilfrage:

-Wie lässt sich die Abspeicherung eines Worteintrags in der Terminologie Ihres Datenbanksystems beschreiben?

Teilantwort:

--

Teilfrage:

- Welche Möglichkeiten vermissen Sie bei Ihrem DBMS?

Teilantwort:

--

9. FRAGE:

Sind ausser der codierten Lexikoninformation noch andere Informationen codiert? (Affixlisten/ andere)

ANTWORT:

--

10. FRAGE:

In welcher Programmiersprache operieren Sie über dem Wörterbuch?

ANTWORT:

PL1

11. FRAGE:

Mit welchen Zugriffsalgorithmen operieren Sie? (VSAM/ hash/ B-Baum/ andere)

ANTWORT:

ISAM

12. FRAGE:

Welche Felder können als Zugriffsschlüssel verwendet werden? (nur ein Schlüssel/ alle Felder/ andere)

ANTWORT:

Für den Direktzugriff dient die Wortform als Schlüssel.

13. FRAGE:

Wie gross ist Ihr Wörterbuch? (Anzahl Einträge/ Anzahl Grundformen/ Anzahl abgeleitete Formen/ andere Grösßenmerkmale)

ANTWORT:

300'000

14. FRAGE:

Maximale Wortlänge?

ANTWORT:

40 Zeichen

15. FRAGE: Maximale Anzahl von Informationskategorien (Codes) pro Lexem?

ANTWORT:

ist wortklassenabhängig, für Verben ca 40

16. FRAGE:

Besteht die Möglichkeit, Einträge zu verknüpfen? (Verknüpfung Lemma- Formen/ Thesaurusrelationen/ Frames/ andere)

ANTWORT:

nur durch expliziten Zugriff auf verschiedene Lexikoneinträge

17. FRAGE:

Welche LDV-Programme (oder Systeme) benutzen Ihre Wörterbücher?

ANTWORT:

Da als Datenbank konzipiert, sind z.Zt. keine LDV-Systeme
direkt an das Lexikon angeschlossen.

18. FRAGE:

existiert eine Schnittstelle zur DIN 2341 (z.B. MATER)? [**7]

ANTWORT:

z. Zt. nicht

19. FRAGE:

Welche konventionellen Wörterbücher dienten als Grundlage? (für den
Wortschatz/ für die Codierung/ andere Informationen)

ANTWORT:

vgl. SDV 2/1980 s. 164 ff

20. FRAGE:

Ist es Ihnen möglich, einen typischen Auszug aus dem Wörterbuch und
eine Erklärung dews codes beizulegen?

ANTWORT:

Vergl. Hinweise nach dem Fragebogen.

21. FRAGE:

Literaturhinweise:

ANTWORT:

Maschinenlesbare deutsche Wörterbücher, Dokumentation, Vergleich,
Integration. Tübingen 1983

Hinweise.

"Die Fragen 4., 7. und 20. betreffen den Datenbankaufbau, der nur
in grösserem Ausmass darstellbar ist. Eine solche Darstellung ist
z.Zt. in Form eines Projektberichts in Arbeit."

Beigelegte Literatur.

- Jan Brustkern, W. Schulze, Aufbau einer Wortdatenbank für das
Deutsche. (Brustkern83)

Später zugestellt wurde:

- Gerlind Heinze, Jan Brustkern, Das Lexikonsystem ALEXSYS
(Heinze85)

*

Die Eintragsstruktur des Lexikonsystems wurde aus der Analyse der Teillexika gewonnen. In (Hess83) wird im Detail darauf eingegangen. In (Brustkern83) sind die aus der Analyse gewonnenen Informationseinheiten in die 7 folgenden Klassen zusammengefasst:

7 Informationsklassen

1. Worteintrag (WE)

Der Worteintrag gliedert sich in die beiden Kategorien "Wortform" (WF) (graphische Repräsentation des Wortlauts) und "Versionen" (V); hier werden Angaben zu graphischer Silbentrennung und Schreibvarianten gegeben.

2. Phonetik/Phonologie

Hierunter werden phonetisch/phonologische Transkriptionen der Wortform sowie Angaben zur Prosodie subsumiert.

3. Wortklasse

Die Angabe einer Wortklasse erfolgt nach einer festgelegten Wortartenklassifikation. Das Wortklassensystem des kumulierten Lexikons wird in 3.2.2 vorgestellt.

4. Morphologie

Hier sind Angaben zur Flexion und/oder zur Wortbildung der Wortform zusammengefasst. Hinzutreten kann die Angabe eines Lemmanamens.⁴⁾

5. Syntaktische Angaben

Angaben zur Oberflächen- und Tiefenstruktur nach verschiedenen Grammatikmodellen.

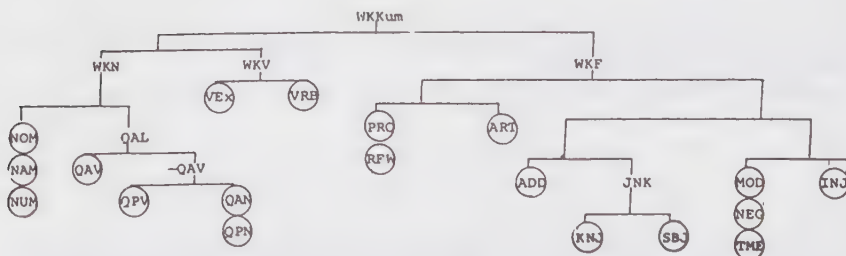
6. Semantische Angaben

Angaben zur Beschreibung und Klassifizierung der Bedeutung der Wortform, wie z.B. 'semantic markers', semantische Relationen, semantische Umgebung, semantische Definition u.ä..

7. Pragmatischer Kommentar

Spezifikation zum Gebrauch der Wortform. Dies sind u.a.:

- Belegstellenangaben
- Beleghäufigkeiten
- Fachgebietskennzeichen



Hierarchie der Wortklassen im kumulierten Lexikon

Figure 15. 'Informationsklassen (ALEXSYS)': (Brustkern83, p. 165)

*

Auch das Wortklassenschema wurde aufgrund der Daten in den Einzelllexika erstellt.

"Es kann nicht das Ziel eines auf dem Prinzip der Kumulation beruhenden Verfahrens sein, ein eigenständiges, einer spezifischen linguistischen Theorie folgendes Klassifikationssystem der Wortklassen des Deutschen zu schaffen.

Folgende Wortklassen werden anhand der Daten der Einzellexika für das kumulierte Lexikon zugelassen (Sie sind grösstenteils syntaxorientiert)":
(Brustkern83, p. 172)

1. NOM (NOMEN)

Hierunter fallen alle Substantive, substantivierte Adjektive und Infinitive. Aus praktisch/technischen Gründen sind NOMINA PROPRIA nicht eingeschlossen.

2. NAM (NOMEN PROPRIUM)

Diese Wortklasse wurde konstituiert, um den nominalen Bereich nicht ausufern zu lassen. Wortformen der Klasse NAM sind generell als spezifisch markierte Nomina aufzufassen:
NAM:= NOM + sem..

3. NUM (NUMERALIA)

Hierzu werden Kardinalia und anderer funktionsspezifische NUMERALIA gezählt, nicht aber Ordinalia, da diese nominal fungieren.

4. QAL (QUALIFIER)

Unter QAL wird die Bezeichnung einer Supraklasse verstanden, die mehrere WKK zusammenfaßt. Sie bezeichnet diejenige Menge von WKK, deren Elemente zur irgendwie gearteten Qualifizierung anderer Wortformen im Syntagma dienen. Die involvierten WKK können nach syntaktisch-funktionalen Gesichtspunkten beschrieben werden:

4.1 QAV (QAL, ADVERB)

QAL, die syntaktisch nur die Position eines Adverbs einnehmen können (also nicht prädikativ/attributive Alternanten besitzen); sie sind flexionsmorphologisch unmarkiert:

4.2 QPV (QAL, PRÄDIKATIV/ADVERB)

QAL, die sowohl prädikativ als auch adverbial fungieren. Prädikativ sind sie $\emptyset$ -markiert, adverbial unmarkiert.

4.3 QAN (QAL, ADNOMINAL)

Attributive Alternanten der QPV, flexionsmorphologisch kasus/genus-markiert; erscheinen ausschließlich in Vollformen-Lexika.

4.4 QPN (QAL, PRÄDIKATIV/ADNOMINAL)

Eine kleine Gruppe von QAL, die stets markiert nicht als ADVERB fungieren können. Sie besitzen z.T. kein Performanzlemma.

-
5. PRO (PRONOMINA)
- Hierunter werden sämtliche nominale und Satzsubstitutiva verstanden; ihre transformatorische Initiative wird regelhaft in der Informationsklasse 'Syntax' beschrieben (z.B. Relation, Negation, Frage).
6. ART (ARTIKEL)
- WK-bezüglich wird nicht zwischen determinierten und indeterminierten Artikel unterschieden (die Differenzierung erfolgt durch die Beschreibungsart des Lemmanamens).
7. RFX (REFLEXIVES FORMWORT)
- Die WK ist auf Pronomina in reflexiver Funktion begrenzt. Sie unterscheiden sich von den Akkusativa der entsprechenden Personalpronomina durch den andersartigen Lemmanamen.
8. ADD (ADDITIVE)
- Präpositionen und Postpositionen sowie Bestandteile prä/postpositionaler Komplexe.
9. JNK (JUNKTIVE)
- Supraklasse für die WKK KNJ und SBJ, im Lexikon nicht realisiert.
- 9.1 KNJ (KONJUNKTION)
- 9.2 SBJ (SUBJUNKTION)
10. MOD (MODALES FORMWORT)
- Die WK beinhaltet sämtliche Satzmodifikatoren mit Ausnahme der Negationspartikeln.
11. NEG (NEGATION)
12. VRB (VERB)
- Sowohl finite als auch infinite (der verbalen Rektion folgende) Verbalformen werden in der WK VRB zusammengefaßt.
13. VEX (funktionspezifische Verben)
- Supraklasse für Modal- und Hilfsverben:
- 13.1 VEH (haben)
- 13.2 VEB (sein)
- 13.3 VEW (werden)
- 13.4 VEM (Modalverben)
14. TME (TMESISFÄHIGES PRÄVERB)
- WK aller Präverbien, die in Tmesis stehen können.
15. INJ (INTERJEKTIONEN)

Figure 16. 'Wortklassen (ALEXSYS)': (Brustkern83, p.172)

2.2.6 Das Lexikon TEAM.

TEAM ist eine multilinguale Terminologiedatenbank (Deutsch, Englisch, Französisch, Spanisch, Russisch, Italienisch, Portugiesisch, Niederländisch).

Das Projekt.

Die technische Terminologiedatenbank TEAM (Terminology Evaluation And Management) ist ein Programmsystem, das durch den Sprachendienst der Firma Siemens entwickelt wurde und laufend erweitert wird.

Für den Benutzer (in erster Linie technische ÜbersetzerInnen, TerminologInnen und NormenspezialistInnen) bietet das System TEAM Verarbeitungs- und Auskunftsmöglichkeiten, die

"vom automatischen Satz aus der Datenbank abgeleiteter Fachwörterbücher und der Ausgabe selektierter Spezialglossare als Hilfsmittel für umfangreiche Übersetzungsaufgaben bis zur Dialogabfrage des gesamten mehrsprachigen Terminologiebestandes" (Schulz81, p.87)

reichen.

Grundsätzlich werden drei Abfragemethoden unterschieden:

1. Einzelabfrage von Terminologie-Einträgen
2. Systematische Abfragen
3. Textbezogene Abfragen

Die Einzelabfrage gleicht dem Nachschlagen in einem herkömmlichen gedruckten Wörterbuch, allerdings mit einigen zusätzlichen Vorteilen für den Benutzer:

- sehr grosser Suchraum (Mehrere Millionen Terme).
- neuester Stand (vorausgesetzt die Terminologiedatenbank wird gut gepflegt).
- Synonyme.
- Output wahlweise auf Schirm oder gedruckt.
- Abfragen über diverse Indices.

2. "Im Gegensatz zur Einzelabfrage, bei der - wie in einem Wörterbuch - bestimmte Fachausdrücke, d.h. einzelne Wörter oder Wortgruppen gesucht werden, ist das Ziel einer systematischen Abfrage eine bestimmten Kriterien genügende Untermenge aller in der Datenbank gespeicherten Einträge." (Schulz81, p.95).

Da in der mehrsprachigen Terminologiedatenbank TEAM die einzelnen terminologischen Einträge in verschiedenen Sprachen verzeichnet und

den einzelnen Einträgen über ihre Kategorien Fachgebietsskennungen zugeteilt sind, ist es möglich, mit geeigneten Selektionskriterien beliebige technische Fachwörterbücher mit wählbaren Sprachpaaren und passenden Indexteilen zu generieren. Ein sehr schönes Muster zeigt folgende Figur:

Seitenzähler <i>m</i> COBOL / счетчик страниц / page counter	Zugriffsmethode <i>Sw</i> / метод индексно-последовательного доступа / indexed sequential access method (ISAM)
Sektor <i>m</i> Mpl. <i>Sw</i> / сектор <i>n</i> / sector <i>n</i>	sequentieller Zugriff / последовательное обращение.
Sektoradresse <i>f</i> ProzRe. <i>Sw</i> / адрес сектора / sector address	sequentieller Zugriff / последовательный доступ / sequential access
Sekundärpeicher <i>m</i> DigDV / память второго уровня, внешняя память / secondary storage	seriell <i>adj</i> <i>Sw</i> / последовательный <i>adj</i> , серийный <i>adj</i> / serial <i>adj</i>
selbstdefinierender Wert <i>Sw</i> / самоопределяющаяся величина / self-defining value, self-defining term	Serienbetrieb <i>m</i> DigDV, DU / последовательный режим работы / serial operation
selbstkorrigierender Code <i>Sich</i> , DU / корректирующий код / error-correcting code, error-correction code, self-correcting code	Serien-Parallelbetrieb <i>m</i> DU / последовательно-параллельный режим / series parallel operation
selbstladend <i>adj</i> <i>Sw</i> / самозагружающийся / self-loading <i>adj</i>	Serien-Parallel-Digitalrechner <i>m</i> DigDV / цифровая вычислительная машина параллельно-последовательного действия / parallel series computer
selbstladender Speicherausgang <i>Sw</i> / самозагружающаяся распечатка памяти / self-loading memory print (SLMP)	Serien-Parallelsystem <i>n</i> DU / последовательно-параллельная система / series parallel system
Selbstorganisation <i>f</i> Kyb / самоорганизация / self-organization <i>n</i>	Serien-Parallelumsetzung <i>f</i> DU / последовательно-параллельное преобразование / series parallel conversion
selbstprüfender Code <i>Sich</i> , DU / код, обнаруживающий ошибки / error-detecting code, self-checking code	Serienrechner <i>m</i> DigDV / последовательная ЭВМ, цифровая вычислительная машина последовательного действия, машина последовательного действия / serial computer
Selektion <i>f</i> Lk / выбор <i>m</i> , селекция / selection <i>n</i>	Seriensystem <i>n</i> DU / последовательная система / serial operation
Selektor <i>m</i> Lk / селектор <i>m</i> / selector <i>n</i>	Servomechanismus <i>m</i> DigDV / сервоустройство <i>m</i> / servomechanism <i>n</i> , servo <i>n</i>
Selektorkanal <i>m</i> ZE / селекторный канал / selector channel	Servosystem <i>n</i> ProzRe, NumSt / сервосистема / servo system
Semantik <i>f</i> MaÜbers / семантика / semantics <i>pl</i>	Setzen Programmaske <i>Befehl</i> / ввод программной маски / Set Program Mask; - Speicherschlüssel <i>Befehl</i> / ввод ключа защиты памяти / Set Storage Key
semantische Information <i>MaÜbers</i> , Kyb / семантическая информация / semantical information, semantical information content, amount of semantic information	Setzmaschine <i>f</i> Satz, Info / наборная машина / typesetting machine
semantischer Gehalt eines Zeichens Kyb / значение <i>n</i> / meaning <i>n</i> ; - Informationsbetrag <i>MaÜbers</i> , Kyb / семантическая информация / semantical information, semantical information content, amount of semantic information; - Informationsgehalt <i>MaÜbers</i> , Kyb / семантическая информация / semantical information, semantical information content, amount of semantic information	Sheffer-Funktion <i>f</i> Boole / штрих Шеффера, функция Шеффера / Sheffer's stroke, Sheffer function, Sheffer stroke function
Semiotik <i>f</i> MaÜbers, Kyb / семиотика / semiotics <i>pl</i>	Sheffersche Funktion <i>Boole</i> / штрих Шеффера, функция Шеффера / Sheffer's stroke, Sheffer function.
sequentiell, indizierte -e	

Figure 17. 'Datenbankgeneriertes technisches Wörterbuch': mit nicht-lateinischem Alphabet (System TEAM)

"Terms in languages using non-Roman alphabets are transliterated

into Roman characters when they are recorded. Since this is done in accordance with the fully reversible transliteration scheme developed by ISO, the terms can be reconverted and output by the computer in their original form, provided the output unit is equipped with the proper character set."
(Vollinhals82, p162)

3. Textbezogene Abfragen beruhen auf der Integration von Textverarbeitung und Datenbankkonsultation. So wie es bei der wörterbuchunterstützten Rechtschreibkontrolle darum geht, jedes Textwort mittels Lexikonabgleich auf Orthographie zu prüfen, kann ein Übersetzer jeden Fachterm automatisch abgleichen lassen und interlinear denselben Term in der Zielsprache in den zu übersetzenden Text einfügen lassen. Das ist nicht mit automatischer Übersetzung zu verwechseln, denn es fehlt jegliche Syntax- oder Semantikanalyse. Die Hilfe für den Übersetzer durch das Interlinear ist aber gerade bei Fachtexten gross, kommt es doch besonders auf die terminologisch richtige Wortwahl an. Die richtige Syntax ist natürlich auch wichtig, aber sie ist nicht das spezifische Merkmal von Fachübersetzungen.

Die Antworten.

1. FRAGE:

Verwenden Sie mehrere Typen von Lexika? (Vollformenwörterbuch/ Grundformenwörterbuch/ Stammwörterbuch/ andere)

ANTWORT:

Grundformenwörterbuch

2. FRAGE:

Welchen Zwecken dient das Lexikon? (Lemmatisierung/ Syntax/ Semantik/ maschinelle Übersetzung/ Information, Dokumentation/ Künstliche Intelligenz/ andere)

ANTWORT:

-maschinelle Übersetzung.
-Terminologiehilfe für Übersetzer.
-Wörterbuchherstellung

3. FRAGE:

Welches Wortklassenschema wird zugrundegelegt?

- Wortarten (Tabelle, ev. Abkürzungen).
- Welche Wortarten unterteilen Sie in welche Unterklassen?
- Welche Wortarten fassen Sie zu welchen Oberklassen zusammen?

ANTWORT:

Siehe TEAM-Kategorienschema unter Wortarterfassung

4. FRAGE:

Welche Informationsfelder werden pro Eintrag codiert?

(zusammengehörende Informationsfelder (Codes) als solche markieren)
Falls die Einträge nach Wortarten differenziert codiert sind, dies
in analogen Tabellen darstellen.

ANTWORT:

siehe TEAM-Kategorienschema

5. FRAGE:

Welchen Rechner benutzen Sie?

ANTWORT:

Siemens 7.5xx (BS2000)

6. FRAGE:

Ist Ihr Verfahren fileorientiert oder datenbankorientiert?

ANTWORT:

fileorientiert, datenbankorientiert. Je nach Verarbeitungszweck

7. FRAGE:

- Aus welchen Feldern ist ein logischer Record aufgebaut?

Teilantwort:

Siehe TEAM-Kategorienschema

- Welche Recordarten benutzen Sie?

Teilantwort:

--

- Entsprechen Ihre Recordarten den Wortarten?

Teilantwort:

nein

8. FRAGE:

Falls datenbankorientiert:

- welchen DB-typ verwenden Sie? (CODASYL/ Baumhierarchie/ Inverted
File/ Relational/ andere)

ANTWORT:

Kombination Alpha-Index / num. Index

Teilfrage:

- Wie lautet die Bezeichnung Ihrer Datenbank als kommerziell
erwerbbares DBMS?

Teilantwort:

TERMAS

Teilfrage:

-Wie lässt sich die Abspeicherung eines Worteintrags in der Terminologie Ihres Datenbanksystems beschreiben?

ANTWORT:

--

Teilfrage:

- Welche Möglichkeiten vermissen Sie bei Ihrem DBMS?

Teilantwort:

--

9. FRAGE:

Sind ausser der codierten Lexikoninformation noch andere Informationen codiert? (Affixlisten/ andere)

ANTWORT:

--

10. FRAGE:

In welcher Programmiersprache operieren Sie über dem Wörterbuch?

ANTWORT:

ASSEMBLER

11. FRAGE:

Mit welchen Zugriffsalgorithmen operieren Sie? (VSAM/ hash/ B-Baum/ andere)

ANTWORT:

VSAM

12. FRAGE:

Welche Felder können als Zugriffsschlüssel verwendet werden? (nur ein Schlüssel/ alle Felder/ andere)

ANTWORT:

Abfrage kann alphabetisch in der gewünschten Sprache oder nach der Eintragsnummer erfolgen.

13. FRAGE:

Wie gross ist Ihr Wörterbuch? (Anzahl Einträge/ Anzahl Grundformen/ Anzahl abgeleitete Formen/ andere Grössenmerkmale)

ANTWORT:

ca 1,5 bis 2.0 Mio Einträge

14. FRAGE:

Maximale Wortlänge?

ANTWORT:

Der terminologische Eintrag kann eine maximale Länge von 4000 Byte haben

15. FRAGE: Maximale Anzahl von Informationskategorien (Codes) pro Lexem?

ANTWORT:

98 (vgl. TEAM-Kategorienschema)

16. FRAGE:

Besteht die Möglichkeit, Einträge zu verknüpfen? (Verknüpfung Lemma- Formen/ Thesaurusrelationen/ Frames/ andere

ANTWORT:

Möglichkeit der logischen Verknüpfung der diversen Selektionskriterien

17. FRAGE:

Welche LDV-Programme (oder Systeme) benutzen Ihre Wörterbücher?

ANTWORT:

Programmsystem TEAM

18. FRAGE:

existiert eine Schnittstelle zur DIN 2341 (z.B. MATER)? [**7]

ANTWORT:

ist derzeit in Arbeit

19. FRAGE:

Welche konventionellen Wörterbücher dienen als Grundlage? (für den Wortschatz/ für die Codierung/ andere Informationen)

ANTWORT:

--

20. FRAGE:

Ist es Ihnen möglich, einen typischen Auszug aus dem Wörterbuch und eine Erklärung des Codes beizulegen?

ANTWORT:

siehe TEAM-Kategorienschema

21. FRAGE:

Literaturhinweise

- Meta No 2/1982
- Philips terminology bulletin No 1/2, August 1979
- Babel No 2/1981

Hinweise.

Beigelegtes Kategorienschema

TEAM-Eintragsschema*) - Kurzfassung -	Stand	Verfasser
	08/1983	Hohnhold
Übersicht		
	Blatt - 1 -	
00-09	Allgemeiner Teil des terminologischen Eintrags (auf Blatt 2 im einzelnen dargestellt) (= Eintragskopf)	
10-89	Sprachenteil des terminologischen Eintrags, unterteilt in die Sprachenblocks	
10-19	Deutsch (auf Blatt 2 im einzelnen dargestellt)	
20-29	Englisch	
30-39	Französisch	
40-49	Spanisch	
50-59	Russisch	
60-69	Italienisch	
70-79	Portugiesisch	
80-89	Niederländisch	
90-98	(nicht fest zugeordnet)	
99	Eintragsendefeld	

Figure 18. 'Kategorienschema (TEAM)': Übersicht

Das Kategorienschema besteht im wesentlichen aus dem Eintragskopf (00-09) und den Sprachenteilen (für jede Sprache 10 Kategorien). Die folgende Figur zeigt die Felderbelegung für den Kopf und für einen Sprachenteil im einzelnen:

00	Eintragsadresse.
01	(frei)
02	(frei)
03	Qualitätskennung des (mehrsprachigen) Eintrags, d. h. Bewertung der Äquivalenzbeziehungen; bei einsprachigen Einträgen ersatzweise Angabe der Einsprachig- keit. B e i s p i e l e : E = Einsprachiger Eintrag A = Äquivalenz nicht gesichert bzw. Äquivalenzgrad nicht definiert; "Arbeitsbenennung" d = Äquivalenz gesichert bzw. Äquivalenzgrad definiert; endgültig geklärte Benennung
04	Zeitpunkt der Erarbeitung oder Eingabe des Eintrags, oder Zeitpunkt einer gezielten Überprüfung, Aktualisierung oder Erweiterung.
05	Teilbestandskennung.
06	Fachgebietskennung(en).
07	(frei)
08	(frei)
09	Geltungsbereichskennung(en).
10	Benennung oder fachsprachliche Wendung, ggf. mit Abkürzung in *() und/oder mit klärendem Zusatz in () oder <u>I</u> ⁻ <u>7</u> . Grundform. Mehrwortbenennungen und Wendungen in natürlicher Gliedfolge. Natürliche Schreibweise. Wortart der Benennung in Feld 10. Quelle(n) der Benennung in Feld 10. (frei) Definition(en) oder definitionsähnliche Erläuterung(en), jeweils mit Quelle; oder Hinweis auf eine Definition in der Literatur. Anwendungsbeispiel(e) der Benennung in Feld 10 (= Kontext(e)), jeweils mit Quelle. Synonym(e) der Benennung in Feld 10, jeweils mit Quelle und Wortart; ggf. mit Abkürzung und/oder klärendem Zusatz wie für Feld 10 beschrieben. Suchwort (-wörter) zur Benennung bzw. Wendung in Feld 10. (frei) (frei)
20-29	(analog)
etc.	

Figure 19. 'Kategorienschema (TEAM)'

TEIL 3 VERGLEICHENDE AUSWERTUNG.

Bei diesem Vergleich werden in den Kommentaren auch die den Bogen beigelegten Arbeitsberichte und andere Publikationen zu den jeweiligen Projekten berücksichtigt. Dies nicht zuletzt deshalb, weil die Fragen teilweise so formuliert sind, dass sich Literaturhinweise als Antwort anbieten:

In diese Kategorie fällt z.B. die Frage Nr. 17.

"Welche LDV-Programme (oder Systeme) benutzen Ihre Wörterbücher?"

Dazu gehört auch die Frage Nr. 21: "Literaturhinweise".

Der Vergleich erschöpft sich deshalb nicht in tabellarischen Gegenüberstellungen der Antworten, sondern enthält neben solchen Überblicken auch Hinweise auf fehlende Antworten oder auf zusätzliches Material. Zusätzlich werden die Antworten kommentiert, wo dies erforderlich ist.

3.1 * INHALTLICHE GRUPPIERUNG DER ANTWORTEN.

Die Reihenfolge der Fragen mit ihren synoptischen Antworten unterscheidet sich von der ursprünglichen Reihenfolge in der für den Versand bestimmten Fassung. In der versandten Fassung wurden die Fragen nicht nach thematisch-systematischen Gesichtspunkten angeordnet. Der Zweck einer unsystematischen Anordnung bestand darin, die einzelne Frage in den Vordergrund zu stellen, um bei der Antwort die Gewissheit zu haben, dass die Antworten nicht vom Fragebogenkontext beeinflusst werden. Aber um das Ausfüllen des Bogens zu erleichtern, wurden Abhängigkeiten explizit erwähnt. (Dies betrifft die Fragen Nr. 6, 7, 8 und deren Teilfragen). Zur Sicherheit wurden die Fragen 20 und 21 eingebaut, um auch Informationen zu erhalten, nach denen nicht explizit gefragt wurde, oder um eine Alternative anzubieten, wenn das Ausfüllen des Fragebogens als eine zu schematische Einengung empfunden werden sollte.

Die Fragen wurden teilweise im multiple-choice-Verfahren gestellt und teilweise offen gehalten. Aber auch bei den multiple-choice-Fragen wurde immer eine Ausweichmöglichkeit angeboten, damit durch die Formulierung der Fragestellung die Antworten nicht präjudiziert wurden.

Um den Vergleich der Antworten übersichtlicher zu gestalten, wurden die 21 Fragen in 6 thematische Gruppen aufgeteilt. Der Zweck dieser Umgruppierung besteht darin, Bezüge zwischen den Fragen sachlogisch evident zu machen und bei der Präsentation der Antworten einen

textuellen Zusammenhang anstelle isolierter Antwort-Tabellen zu setzen.

Die Gruppen sind:

1. Fragen auf der linguistischen und der konzeptionellen Ebene. Zu dieser Gruppe gehören:

- a) Frage Nr.2, die nach den hauptsächlichsten funktionalen Zwecken des Lexikons fragt. Sie könnte allenfalls auch noch der 5. thematischen Gruppe zugeordnet werden.
- b) Frage Nr.1, welche nach morphologischen Kriterien den oder die Lexikontypen erfragt.
- c) Frage Nr.3: Wortklassenschema.
- d) Frage Nr.4: Struktur des Eintrags.
- e) Frage Nr.16, bei der es um die möglichen Relationen zwischen den Einträgen geht. Sie könnte z.T. auch der 3. thematischen Gruppe (Implementation) zugeordnet werden.
- f) Frage Nr.20. (Code-Erklärung) reicht ebenfalls teilweise in die 3. Gruppe hinein.

Diese 6 Fragen werden beim synoptischen Vergleich als Einheit betrachtet und auch entsprechend kommentiert.

2. Fragen zur Datengrundlage der Lexika. Dazu sind zu rechnen:

- a) Frage Nr.9, wo nach zusätzlichen lexikalischen Informationsquellen gefragt wird; sie hat auch einen Bezug zur 1. Gruppe.
- b) Frage Nr.19: Vorlagen aus (konventionellen) Wörterbüchern.

3. Fragen zur Implementation. [**1] Dazu gehören:

- a) Frage Nr.6: Filesystem oder Datenbank.
- b) Frage Nr.7: Recordarten und Recordaufbau.
- c) Frage Nr.8: DB-typ und DBMS.
(Die Fragen sieben und acht sind alternativ.)
- d) Frage Nr.11: Zugriffsalgorithmus.
- e) Frage Nr.12: Schlüssel.
- f) Frage Nr.10: Programmiersprache.

4. Fragen nach quantitativen Merkmalen oder Restriktionen. Dies sind:

- a) Frage Nr.13: Momentane Grösse des Lexikons.
- b) Frage Nr.14: Wortlängenrestriktion.
- c) Frage Nr.15: Eintragslängenrestriktion.

5. Fragen nach dem Umfeld. (System, Projekt, Schnittstellen). Dazu gehören:

- a) Frage Nr.17: Lexikonbenutzung.
- b) Frage Nr.18 Normierte Schnittstellen.

6. Hardware.

betrifft nur die Frage Nr.5 (Rechnertyp).

Übrig bleibt die Frage Nr.21. Sie lässt sich keiner Gruppe zuordnen.

*

Die folgenden Synopsen sind nach diesen 6 Gruppen geordnet, und die Kommentierung erfolgt z.T. gruppenweise und z.T. direkt unter der betreffenden Synopse. Falls der tabellarische Überblick für sich selber spricht, wird auf einen Kommentar verzichtet.

3.1.1 Linguistische Konzeption.

Die Antworten auf die in dieser Gruppe zusammengefassten Themen

- Applikationszusammenhang ("Zweck"),
- Morphologische Qualität des Eintragsheaders (Lemma) als Kriterium für den Lexikontyp,
- Wortklassenschema,
- Eintragsstruktur ("Informationsfelder pro Eintrag"),
- Verflechtungsstruktur ("Verknüpfungen zwischen den Einträgen"),
- Code-Erklärungen

erlauben Rückschlüsse auf die linguistische Konzeption des Lexikonsystems. Die Reihenfolge der Fragen innerhalb dieser Gruppe ist so beschaffen, dass von "handfesteren" zu komplexeren Fragen

vorangeschritten wird.

Obwohl z.B. die Frage nach "Informationsfeldern pro Eintrag" (Frage Nr. 4) für eine primäre Charakterisierung des Lexikons geeigneter zu sein scheint (denn sie deutet auf das Abfolgeschema hin) als die der Gruppe vorangestellte Frage Nr. 2 nach dem hauptsächlich funktionalen Zweck des Lexikons, so ist es methodisch doch klarer, zuerst den Verwendungszweck des Lexikons anzugeben, bevor dessen Komponenten präsentiert werden. Ebenso ist es kognitiv adäquater, zuerst zu erfahren, ob das Lexikon ein Vollformenlexikon oder ein Stammwörterbuch ist, bevor die Verflechtungsstruktur abgehandelt wird. Das Gleiche gilt für das Wortklassenschema und die Code-Erklärungen.

[a]

2. FRAGE:

Welchen Zwecken dient das Lexikon? (Lemmatisierung/ Syntax/ Semantik/ maschinelle Übersetzung/ Information, Dokumentation/ Künstliche Intelligenz/ andere)

Diese Frage zielt darauf ab, das entsprechende Lexikon grob in seinem linguistischen Applikationszusammenhang zu charakterisieren. Die in den Antworten verwendeten Abkürzungen beziehen sich auf die Auswahlmöglichkeiten in der Frage. Das unter SADAW stehende Akronym CTX bedeutet "Computergestützte Texterschließung" und meint den Applikationszusammenhang der maschinellen Indexierung zur Fachinformation mit Hilfe des Wörterbuchs SADAW. (vergl. Zimmermann83, p7)

ANTWORTEN:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
LEM. SYN. K.I.	SYN. SEM. K.I.	SYN. SEM. K.I. PHONETIK	SYN. SEM. M.Ü. DOKU. (CTX)	SYN. SEM. M.Ü. DOKU. K.I. ALLG. WORDATA- TENBANK	M.Ü. TERM. WÖRTER- BUCHHER- STELLUNG

Die Antworten zeigen, dass sich die Systeme von ihrem Anwendungszusammenhang her betrachtet durchaus vergleichen lassen, weil keines aus dem in der Frage vorgegebenen Vergleichsrahmen herausfällt. Aus den Antworten lassen sich aufgrund ähnlicher Antwortbilder Gruppierungen ablesen. In Bezug auf die ersten beiden Zeilen der Tabelle herrscht eine auffällige Übereinstimmung zwischen den ersten 5 Systemen. In Bezug auf die dritte Zeile können die ersten 3 Lexikonsysteme als zusammengehörig betrachtet werden. Das 5. System, in dem K.I. auch erwähnt ist, könnte vielleicht auch dazugezählt werden, stellt aber in Bezug auf "Verwendungszweck" einen Sonderfall dar, da die kumulierte Wortdatenbank bewusst nicht auf einen bestimmten Zweck fixiert wurde. In Bezug auf den Anwendungszusammenhang "Maschinelle Übersetzung" herrscht Übereinstimmung zwischen dem 4., dem 5. und dem 6. Lexikon. Phonetik und Terminologie-Wörterbuchherstellung

können als eher spezielle Applikationszusammenhänge gewertet werden. Und noch eine weitere Beobachtung kann anhand der Tabelle gemacht werden: Bezüglich der Vielfalt der Anwendungszwecke herrscht Übereinstimmung zwischen den Systemen SADAW und ALEXSYS. Die auffälligste Spezialisierung bezüglich Applikation kann bei TEAM festgestellt werden.

[b]

1. FRAGE:

Verwenden Sie mehrere Typen von Lexika? (Vollformenwörterbuch/ Grundformenwörterbuch/ Stammwörterbuch/ andere)

Diese Frage entspricht einer gängigen Typisierung maschinenlesbarer Lexika und erlaubt Rückschlüsse auf den effektiven Umfang des Wörterbuchs und auf zusätzliche lexikalische Wissensquellen. Ein Vollformenlexikon wird schnell sehr umfangreich, hat aber den Vorteil, dass flexionsmorphologische Regeln z.T. überflüssig werden. Ein Stammformenlexikon stellt die am wenigsten redundante Speicherung lexikalischer Information dar (wenn man von der theoretisch denkbaren noch radikaleren Komprimierung in Form eines Morphem-Lexikons absieht), ist aber auf einen Regelapparat zur Generierung und Erkennung abgeleiteter und flektierter Formen angewiesen.

In der traditionellen Lexikographie lässt sich der Stellenwert der eben geschilderten Typisierung mit den ein traditionelles Lexikon charakterisierenden Begriffen "semasiologisch" und "onomasiologisch" vergleichen. Diese Begriffe sind aber für maschinelle Lexika weniger charakteristisch, weil sie auf die Anordnung und die Sortierung der Einträge Bezug nehmen, welche in maschinenlesbaren Systemen rasch verändert werden können.

ANTWORTEN:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
GF	SF	VF	GF	VF/GF	VF

(SF=Stammformen, GF=Grundformen, VF=Vollformen)

Der Begriff "Stammform" im Lexikon KIT entspricht nicht dem aus der historisch-vergleichenden Sprachforschung bekannten "Stamm"-Begriff. Eine klare Definition des Terms "Stammform" im Kontext des Lexikons KIT findet sich im Kommentar zur Frage [d] (= 4. Frage des Bogens) innerhalb der hier behandelten ersten Fragegruppe (Linguistische Konzeption).

Die doppelte Antwort unter ALEXSYS bedeutet, dass in diesem Lexikon sowohl Vollformen als auch Grundformen gespeichert sind.

Der Begriff "Grundform", wie er im Kontext des Systems SADAW benutzt wird, charakterisiert H. Zimmermann wie folgt:

"Ziel war es, Wortformen in beliebigen deutschsprachigen Texten mittels eines maschinellen Lexikons und eines darauf bezogenen Algorithmus auf mögliche Grundformen zurückzuführen. Hierzu wurden gemäß der traditionellen Grammatik Wortklassen differenziert,

wobei die Wortformen der sog. Funktionswortklassen (Konjunktionen, Präpositionen...) sowie der reinen Adverbien (z.B. HEUTE, OFT) als solche lexikalisch erfasst wurden, während die Substantive, Verben und Adjektive mit ihrem Stamm (genauer: bei unregelmässiger Wortbildung mit mehreren Stämmen) verzeichnet wurden. Beispiele für derartige Stämme sind HAUS (für HAUS, HAUSE, HAUSES) und HAUSER (für HAUSER, HAUSERN), GEH (für GEH, GEHE, GEHEN, GEHT, GEHTST), GING (für GING, GINGST, GINGE, GINGEN) usf. Diese Stämme erhielten Markierungen bezüglich der möglichen Flexionsendungen und ggf. einen Verweis auf die "Grundform". (Zimmermann83, p. 7)

Die Grundform gilt als 'Repräsentant' für das ganze Paradigma, wobei es willkürlich ist, z.B. für das oben erwähnte HAUS/HAUSER-Paradigma den Nominativ Singular (HAUS) als Grundform zu deklarieren.

[c]

3. FRAGE:

Welches Wortklassenschema wird zugrundegelegt?

- Wortarten (Tabelle, ev. Abkürzungen).
- Welche Wortarten unterteilen Sie in welche Unterklassen?
- Welche Wortarten fassen Sie zu welchen Oberklassen zusammen?

Die Beantwortung der Frage nach den Wortarten im Lexikon erlaubt Rückschlüsse auf das im entsprechenden System verwendete Grammatikmodell, weil die Wortklassen für die Konstituentenstrukturen die präterminalen Symbole bilden. Die Wortartenteilung ist auch deshalb von Interesse, weil viele der im Lexikon verzeichneten morphosyntaktischen und semantischen Angaben wortklassenabhängig sind. Die Frage nach den Wortarten ist keineswegs trivial, da auch in traditionellen Grammatiken unterschiedliche Auffassungen bestehen. Einen systematischen Überblick über die Wortartenproblematik gibt (Stepanova81).

ANTWORTEN:

Weil diese Frage, zusammen mit ihren Teilfragen, relativ umfangreiche Antworten ergeben hat, wird synoptisch nur die Grobklassifikation gezeigt, für die Feinklassierung muss auf grössere Tabellen verwiesen werden, die z. T. bei den nach Projekten geordneten Antworten schon aufgeführt wurden.

Grobklassifizierung:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
VERB	NG	ADJ	VERB	WKN	VERB
NOM	VG	ADJU	ADJ	WKV	ADJ
NOM	VG	ADJU	ADJ	WKV	ADJ
ADJ	Mod	ADV	SUBST	WKF	SUBST
DETDEF	Konj	AUXH	FUNKTIONS		FKTW
DETIND	Praep	AUXS	WÖRTER		
DETQ	Beif	AUXW			
PRN	SatzZ	DET			
POSSPRN		MVERB			
PRP		NEG			
XXX		N			
		NPR			
		ORDZAHL			
		PRAEP			
		VERB			
		WADV			
		WDET			
		WPRON			
		ZAHLWORT			

KOMMENTAR zu den Wortklassen:

Nach Projekten geordnet wird auf bereits aufgeführte Tabellen zu den Wortklassen verwiesen. Auch die Kommentare sind nach Projekten geordnet.

-HAM:
Die aufgelisteten Abkürzungen für die Klassenbezeichnungen sind in den Tabellen der Figuren 6 und 7 erklärt. In der Synopse sind nur die SUTRA-Wortklassen (SURFACE-TRANSFORMATION) verzeichnet.

*

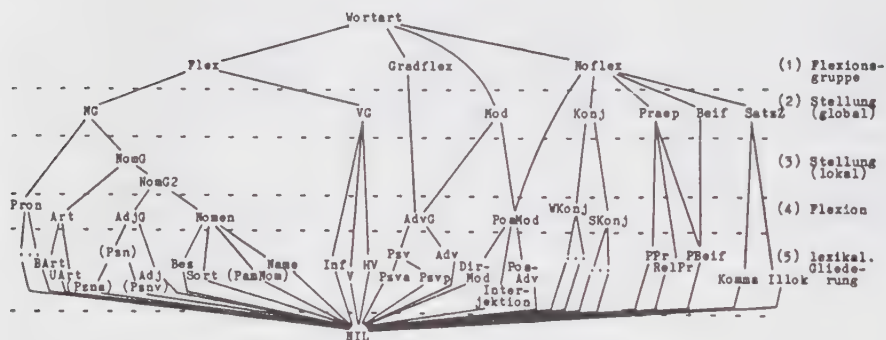
-KIT:
Die Wortarten sind bei KIT als

"unmittelbar operationalisierbare Kontexterwartungen"
(Eimermacher83a, p.5)

in einem Verband geordnet. In dieser Ordnung werden die Wortarten in fünf Ebenen nach verschiedenen Kriterien eingeteilt. Für das Lexikon ist, abgesehen von der in der Synopse unter KIT verzeichneten 3. Ebene (globale Stellung), die 5. Ebene (lexikalische Gliederung) relevant. Der Sinn des Wortartenverbands mit allen seinen 5 Ebenen besteht darin, dass

"Dieser Verband [...] eine erste Operationalisierung von syntaktischen Kontexterwartungen ohne Umweg über (Teil-) Phrasenstrukturen [gestattet]".
(Eimermacher83a, p.7/8)

Der Wortartenverband wird in (Eimermacher83a) auf Seite 6 graphisch dargestellt:



Abkürzungen (mit * gekennzeichnete sind im Text näher erläutert)

Flex = mit Flexion, vs. Gradflex (Steigerung möglich, aber keine weiteren Endungen) bzw. NoFlex
 Geklammerte Wortarten entstehen nur durch Transformation

G = Gruppe: NG ist die Obersorte aller nominalen Wortarten - im Gegensatz zur NP, die den Typ einer Phrase beschreibt

PRONomen	PartiZip	Verb , INfinitiv	SatzKONJunktion
ARTikel	- Nominal (flektiert)	HilfsVerb	WortKONJunktion
(Bestimmt/	- Verbal (unflektiert)	ADVerb	*Präposition als BEIFügung
Unbestimmt)	- präsens Aktiv	*POSitiv MODifikator	*PPr, *RelPr
	- Perfekt (aktiv/passiv)	*DIRekt MOD	SATZzeichen
*BEzeichnung	*PARametrisiertes NOMen		ILLOKution
SORTenbezeichner	NAME		

Figure 20. 'Wortartenverband (KIT)': (Eimermacher83a, p.6)

Die am feinsten aufgegliederte Ebene stellt die 5. Ebene mit ihrer lexikalischen Gliederung dar. Diese Gliederung wird für die Einteilung der Wortstammexperten (WSXP) und der Wortartexperten (WAXP), welche zusammen mit der Endungstabelle sowohl das Lexikon als auch den dezentralen Word-Expert-Parser bilden, benötigt. (Im Rahmen des Kommentars zur nächsten Frage [d] wird auf dieses Konzept näher eingegangen).

Zur Graphik der Wortarten macht M. Eimermacher folgende Anmerkungen, welche die gebrauchten Abkürzungen erklären (Die gebrauchten Beispiele beziehen sich immer auf Wetterberichte, was

aber nicht zum Fehlschluss verleiten sollte, das von Eimermacher vorgestellte Wortklassenschema eigne sich nur für die "Miniwelt" der Wettervorhersage):

"NG's:

- PamNom = 'parametrisiertes' Nomen:

Nomen, das in der SRL-Repräsentation Argumentstellen benötigt. Das sind in der Regel nominalisierte Verben ('die Abschwächung des Hochs').

- Bez = Bezeichnung:

Nomen, das kein festes Individuum benennt (vs. Namen), aber auch nicht als Sortenbezeichner verwendet wird. Beispiele wären Regen und Schnee, wenn als Sorte nur NIEDERSCHLAG verwendet wird. (Wenn man verschiedenen Regenfällen die gemeinsame Sorte REGEN geben will, ist Regen kein BEZ, sondern SORT - die Trennung ist also willkürlich bzw. von der Textklasse beeinflusst).

Verben brauchen für die Textklasse 'Wetterberichte' zunächst nicht weiter gegliedert zu werden.

Modifikatoren:

- PosMod = Positiv-Modifikatoren:

Modifikatoren im Positiv (ohne Steigerungsformen), z.B. 'ZIEMLICH schwach'

- PosAdv = Positiv-Adverbien:

Adverbien im Positiv (ohne Steigerungsformen), z.B. 'heute'

- DirMod = Direkt-Modifikatoren:

Modifikatoren, die sich auf das direkt folgende Wort beziehen

- Interjektionen:

Hier werden nur diejenigen angesprochen, die nicht am Satzanfang (allein) stehen können (*'Nicht komme ich'). 'Dennoch' wird also als PosAdv eingeordnet ('Dennoch komme ich').

Präpositionen:

- PBeif: Präposition als Beifügung ('...greift auf Deutschland ÜBER')

- RelPr: Präposition als Relativsatzeinleitung ('..., MIT der Meeresluft herangeführt wird')

- PPr: normale Präposition

SatzKONjunktionen werden eingeteilt in:

- KSKonj (für Konjunktionalsatz: 'weil'),

- KmpKonj (für Komplementärsatz: 'dass'),

- NKonj (nebenordnende Haupt- oder Nebensatzkonjunktion: 'und')"

(Eimermacher83a, p.7)

*

-EVAR:

Die in der Synopse angegebenen Wortarten sind in (Brietzmann84a) auf der Seite 123 mit folgenden Beispielen erläutert:

. ADJ	(attributives) Adjektiv	("schöner")
. ADJU	(unflektiertes) Adjektiv	("schön")
. ADV	Adverb	("gestern")
. AUXH	Hilfsverb "haben"	
. AUXS	Hilfsverb "sein"	
. AUXW	Hilfsverb "werden"	
. DET	Determinans	("der, ein, kein")
. MVERB	Modalverb	("wollen")
. NEG	Negation	("nicht")
. N	Nomen	("Schalter")
. NPR	Nomen proprium	("München")
. ORDZAHL	Ordinalzahl	("achter")
. PRAEP	Präposition	("nach")
. VERB	(Voll-)Verb	("wissen")
. WADV	Frageadverb	("wann")
. WDET	Fragedeterminans	("welcher")
. WPRON	Fragepronomen	("wer")
. ZAHLWORT	Kardinalzahl	("acht")

Figure 21. 'Wortklassen (EVAR)': (Brietzmann84a, p.123)

Die hier aufgelisteten Wortklassen entsprechen denjenigen in Figur 10, allerdings fehlen in der eben angeführten Liste folgende Wortarten, welche in der Liste in Figur 10 vorhanden sind:

- EPAR
- GRADPAR
- INFSSUBJ
- KONJ
- KOORD
- MADV
- PRAEPPRAEP
- PRON
- REFLPRON
- RELPRON
- SUBJ
- VERBPRAE
- VERPAR

Die Differenz zwischen den beiden Listen erklärt sich dadurch, dass die Liste in Figur 10 die Wortklassen zeigt, welche dem

Syntax-Modul bei der ATN-Analyse bekannt sind und die andere Liste die Wortarten enthält, welche im Semantik-Modul instantiiert und ausgewertet werden können.

Vom syntaktischen Gesichtspunkt her betrachtet, sind die Wortarten auf das Inventar der zugelassenen syntaktischen Konstituententypen bezogen, welche sowohl in (Brietzmann83a,p.123/124) als auch in (Brietzmann84b,p.62) verzeichnet sind.

Vom semantisch-pragmatischen Gesichtspunkt her sind die Wortarten als Spezialisierungen des KL-ONE Konzepts WORT angelegt.

Zusammen mit der syntaktischen

"funktionalen Hierarchie der Konstituententypen"
(Brietzmann84a,p.122)

und den

"semantischen Klassen in ihrer Verträglichkeitshierarchie"
(Brietzmann84a,p.122)

stellen die Wortklassen

"das faktische Wissen des Semantik-Moduls dar".
(Brietzmann84a,p.122)

Dieses Wissen ist gemäss dem verwendeten KL-ONE-ähnlichen Repräsentationsformalismus

"in Form einer Konzept-Hierarchie definiert, die entsprechend in drei Teilbäume unterteilt werden kann: die Spezialisierungen von KONSTITUENTE, von WORT und von CLASS."
(Brietzmann84a,p.123)

(Auf diese Konzepthierarchie wird in (2.3.2) noch eingegangen)

Eine neuere Liste der Wortarten, nach den Einteilungskriterien

- "Wortart MIT Semantik" versus "Wortart OHNE Semantik"

und

- "flektierbare Wortart" versus "NICHTflektierbare Wortart"

findet sich im Anhang von (Ehrlich86):

```

wortart :-
    wortart_mit_semantik | wortart_ohne_semantik

wortart_mit_semantik :-
    Adjektiv | Adverb | Determinans | Frageadverb |
    Fragedeterminans | Fragepronomen |
    Infinitivsubjunktion | Konjunktion |
    Koordination | Modalverb | Nomen |
    Nomen_proprium | Ordinalzahl | Praeposition |
    Pronomen | Reflexivpronomen |
    Subjunktion | unflektiertes_Adjektiv |
    Verb | Vergleichspartikel | Zahlwort

wortart_ohne_semantik :-
    Floskel | Hilfsverb_"haben" | Hilfsverb_"sein" |
    Hilfsverb_"werden" | Infinitivpartikel |
    Negationspartikel | Praepraeposition |
    Relativpronomen | Satzwort | Verbprefix

flektierbare_wortart :-
    Adjektiv | Determinans | Fragedeterminans |
    Fragepronomen | Hilfsverb_"haben" |
    Hilfsverb_"sein" | Hilfsverb_"werden" |
    Modalverb | Nomen | Nomen_proprium | Ordinalzahl |
    Praeposition | Relativpronomen |
    unflektiertes_Adjektiv | Verb

```

Figure 22. 'Forts. Wortklassen (EVAR)':

*

-SADAW:

Die in der Synopse angegebenen Wortklassen wurden im Fragebogen als Antwort auf die entsprechende Teilfrage zu Frage Nr. 3 gegeben und sind deshalb ausdrücklich als Oberklassen zu verstehen.

In den Tabellen der Figuren 12 und 13 sind die Wortklassen differenzierter dargestellt. Der dort verwendete Begriff "Analysewortklasse" ist wie folgt zu verstehen:

"Jeder Flexionsform ist eine Wortklasse zugeordnet, die ihr Verhalten (insbesondere ihre Distribution) in Sätzen (bzw. in Syntagmen) im groben charakterisiert."
(Eggers80, p.168)

Die Analysewortklassen sind hauptsächlich nach syntaktischen Kriterien eingeteilt, und es wird auf traditionelle Bezeichnungen zurückgegriffen.

Beispiele:

Wortform	Analyse-WK.		Lemma-Wk.
LAUFEN	Infinitiv	INF	Verb
LAUFEN	fin. Verb	FIV	Verb
GELAUFEN	Partizip II	PTZ 2	Verb
LAUFEN	sub. Infinit.	SBI	Verb
GELAUFENE	Attr. Adj.	ADJ	Verb
LAUF	Substantiv	SUB	SUB

(Beispiele aus (Eggers80), p.168.)

Den hier wesentlichen Unterschied zwischen "Analyse-Wortklasse" und "Lemma-Wortklasse" verdeutlicht die folgende Definition von

"Lemmawortklasse:

Jedem Lemma ist eindeutig eine Wortklasse zugeordnet, die Verb (VRB), Substantiv (SUB), Adjektiv (ADJ) oder Restklasse (FWK) sein kann."

(Eggers80, p.170)

Die von H.-D. Maas im Fragebogen angegebenen Oberklassen entsprechen also den Lemmawortklassen.

Wie bereits erwähnt, sind die Analysewortklassen nach morphosyntaktischen Kriterien differenziert, wobei aber beachtet werden muss, dass die in (Eggers80) auf Seite 34ff aufgezählten Analysewortklassen noch nicht völlig dem in SADAW verwendeten Wortklassenschema entsprechen.

"Das oberste Kriterium bei der Einteilung der Textwörter in Analysewortklassen ist ihr syntaktisches Verhalten und ihre Distribution. Dabei werden noch nicht alle 'Feinheiten' berücksichtigt; diese schlagen sich in den Subkategorien nieder (vgl. Kodierungsanweisungen, Kap.II)."

(Eggers80, p.34)

In der dem Fragebogen beigelegten Benutzeranleitung zu SADAW mit Eingabebeschreibung und Kodieranweisungen werden beispielsweise die Verben in die folgenden 10 Subtypen eingeteilt:

(Auf diese Subtypen geht auch (Eggers80, p60) ein)

- Vollverb. (Code: 0)
- 'haben'- Paradigma. (Code: 1)
- 'sein'- Paradigma. (Code: 2)
- Modalverb. (Code: 3)
- 'werden'- Paradigma. (Code: 4)
- ACI-Verb. (Code: 5)
- Paradigma 'kommen' und 'gehören'. (Code: 6)
- Paradigma eines Verbs, das 'zu' + Infinitiv bzw. Infinitiv mit eingeschlossenem 'zu' fordert. (Code: 7)
- 'lassen'- Paradigma (Code: 8)
- Paradigma 'gehen', 'helfen', 'üben', 'lehren', 'lernen'. (Code: 9)

Bei dieser Subtypenteilung ist auch berücksichtigt, dass ein Verb
"eventuell zu mehreren Subtypen gehören [kann], z.B.

haben als Hilfsverb:	Er hat gestanden.	Kodierung = 1
haben als Vollverb:	Er hat Mut.	Kodierung = 7
	Er hat zu gehorchen."	

(Benutzeranleitung, p.23)

*

-ALEXSYS:

Die in der Synopse genannten Wortklassen entsprechen den in Figur 16 gezeigten Klassen. In der Hierarchie der Wortklassen werden unter "WKN" im nominalen Bereich sowohl Substantive als auch Qualifier verstanden. Hilfs- und Vollverben fallen unter die Klasse "WKV", wobei die Unterklasse "VEx" die Hilfsverben und die Modalverben umfasst, die beide zusammen als "funktionsspezifische Verben" bezeichnet werden.

Die in Figur 16 gezeigten Klassen bilden das Raster, welches auf die Wortklassenschemata der im kumulierten Lexikon aufgenommenen Einzellexika projiziert wurde. Die Wortklassencodes der Einzellexika sind in (Brustkern83) auf den Seiten 177/78 verzeichnet. Auch in (Hess83) sind diese Codes zu finden.

*

-TEAM:

Bei der Terminologiedatenbank TEAM sind die Wortklassen-Angaben in Feld 11 des in den Figuren 18 und 19 gezeigten Kategorienschemas verzeichnet. Diese Wortklassenangaben beziehen sich auf die in Feld 10 des Schemas verzeichnete Benennung oder fachsprachliche Wendung.

Wortklassenangaben spielen jedoch in einem technischen Fachwörterbuch eine andere Rolle als in den fünf bisher erwähnten Lexikonsystemen:

"Der terminologische Ansatz führt natürlich dazu, dass als Benennung nicht nur einzelne Wörter erfasst (und 'übersetzt', d.h. mit ihren Äquivalenten in anderen Sprachen versehen) werden. In den Fachsprachen sind Mehrwortbenennungen sehr häufig (wenn es im Deutschen aufgrund der besonderen Kompositabildung vielleicht auch nicht so augenfällig ist wie in anderen Sprachen), da zahlreiche Begriffe eindeutig nur mit solchen Wortgruppen zu bezeichnen sind (z.B. Adjektiv und Substantiv im Deutschen, englische Komposita, die in der Regel ja nicht zusammengeschrieben werden usw.).

Eigentliche linguistische Dateneinheiten einer Terminologiesammlung sind also nicht einzelne Wörter, sondern oft recht komplexe Syntagmen."
(Schulz79,1.4)

Schulz weist darauf hin, dass diese Syntagmen in ihrer Mehrzahl aus Substantiven oder zusammengesetzten Nominalausdrücken bestehen, aber sie können auch ganze phraseologische Einheiten bilden. Dabei werden sie aber immer noch als Terme ("Benennung" im Sinne von DIN 2330, vgl. Anmerkung [**1] in Teil 1) behandelt:

"Von nicht zu unterschätzender Bedeutung sind dabei Verbalausdrücke, etwa die Verbindungen von substantivischen Benennungen mit Verben, die unter Umständen erst in dieser Verbindung ihre fachsprachliche Bedeutung erhalten. Es kann sich dabei um einfache Verb-Objekt-Verbindungen handeln ('ein Register laden', 'ein Programm binden', 'einen Schalter setzen') oder auch um ganze Sätze mit finiten Verbformen ('ein Programm steigt mit einer Fehlermeldung aus', 'ein Programm gerät in eine Schleife', 'ein Relais spricht an'). Auch diese phraseologischen Einheiten können zusammen mit ihren fremdsprachigen Äquivalenten in gleicher Weise wie einfache Terminologiegleichungen gespeichert und nach analogen Gesichtspunkten und Kriterien geordnet und abgefragt werden."

(Schulz79,1.5)

Trotz dieser Einschränkungen erfüllen Wortklassenangaben auch in technischen Wörterbüchern bisweilen eine wichtige Aufgabe:

"In der Regel wird nur bei einzelnen Wörtern (Benennungen, die nur aus einem Wort bestehen) die Wortart, bzw. bei Substantiven das Genus angegeben. Das hilft dem Benutzer, zwischen Homographen zu unterscheiden. ('Band m./n.', 'close' v./adj.'). Selbstverständlich wird jedes der Homonyme zusammen mit seinen Zusatzinformationen in einem eigenen Eintrag erfasst."

(Schulz79,1.5)

[d]

4. FRAGE:

Welche Informationsfelder werden pro Eintrag codiert?

(zusammengehörende Informationsfelder (Codes) als solche markieren)
Falls die Einträge nach Wortarten differenziert codiert sind, dies in analogen Tabellen darstellen.

Diese Frage steht in einem direkten Zusammenhang mit dem in Teil 1 der vorliegenden Arbeit entwickelten Begriff "Abfolgeschema". Die "explizit kategorisierten Eintragungselemente" der Abfolgeschemata entsprechen den "Informationsfeldern pro Eintrag", die in dieser Frage angesprochen sind. In der Frage ist auch angedeutet, dass diese Informationsfelder gruppiert werden können und Abhängigkeiten aufweisen können.

ANTWORTEN:

Auch bei diesen Antworten wird wie bei [c] nur die grobe Einteilung synoptisch gezeigt. Auf Feingliederungen wird separat hingewiesen. Es ist ferner zu beachten, dass die in der Synopse linear untereinander gelisteten Kategorien in den einzelnen Lexikonsystemen z.T. nicht linear angeordnet sind (entsprechend den in Teil 1 der Arbeit beschriebenen verschiedenartig strukturierten Typen von Abfolgeschemata).

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
WORTART	WXP	GRAPH.-	IDENTI-	WORTHEIN-	EINTRAGS-
STAMM	WSXP	INF.	FIKATIONS-	TRAG	KOPF
SYNONYM	WAXP	PHON.-	WORTLAUT	PHON.	DEUTSCH
FLEXION		INF.	BEDEU-	WORTART	ENGLISCH
KASUS-		WORTART	TUNGS-	MORPHOL.	FRANZÖS.
RAHMEN		MORPHO-	NUMMER	SYNTAX	SPANISCH
GENUS		SYNTAX	INFOR-	SEMANTIK	RUSSISCH
NOMEN-		SEMANTIK	MATIONEN-	PRAGMAT.	ITALIEN.
KLASSE		NORMAL-	TEIL		PORTUG.
(DEKLIN)		FORM			NIEDERL.
VERB-					
KLASSE					
(KONJUG)					

KOMMENTAR zu den Kategorien:

Die Kommentare sind wie in der vorhergehenden Frage projektweise angebracht, wie auch die Hinweise auf bereits aufgeführte Tabellen.

-HAM:

Die in der Synopse genannten Kategorien sind in Figur 8 erklärt und mit Beispielen erläutert. Die Kategorien beziehen sich auf das Wortlexikon. Um einen vollständigen Überblick über die in HAM-ANS verwendeten lexikalischen Kategorien zu haben, müssen auch das Syntagmenlexikon, das Kasusrahmenlexikon, die Flexionssuffixe, semantisches Netz sowie die Repräsentationssprachen SURF und DEEP berücksichtigt werden (vergl. p.73-74). Die in diesen Wissensquellen verwendeten Kategorien sollen nicht vollständig aufgelistet werden, sondern einige Beispiele mögen genügen:

- Paraphrase von Syntagmen
- Semantische Netzrelationen wie:
 - Oberbegriff
 - Obermenge
 - Hat-als-möglichen-Wert
 - Hat-als-Eigenschaft
 - Hat-als-Teil
 - Teilmenge
- Affixlisten

*

-KIT:

Der in (Eimermacher83a,b,c) beschriebene syntaktisch-semantische PROLOG Word-Expert-Parser greift auf vier lexikalische Informationsquellen zu:

- Endungstabelle für die Deflexion
- Wortstammexperten (WSXP)
- Wortartexperten (WAXP)
- Kongruenztabelle [**2]

In der Synopse ist ausserdem noch WXP vermerkt: Wortexperten. Der Zusammenhang zwischen WSXP, WAXP und WXP ist folgender:

"Jeder WAXP ist eine Transformation der Form"

WSXP o Endung(en) --> WXP

Diese Transformationsformel bedeutet, dass ein Wortexperte (WXP) mittels eines WAXP (Wortartenexperte) aus einem Wortstammexperten (WSXP) und Endungen erzeugt wird.

Dieses Konzept stellt eine Weiterentwicklung des WXP-Lexikons dar, welches in (Eimermacher83c) beschrieben ist. Das WXP-Lexikon ist integrierender Bestandteil der semantischen Parserkomponente. Die Grundidee ist folgende:

"Jedem Wort bzw. Affix [wird] ein Wortexperte (WXP) zugeordnet. Er ermittelt aufgrund eines Entscheidungsnetzes durch Kommunikation mit den übrigen für den Satz zuständigen WXP's die aktuelle Bedeutung seines Wortes" (Eimermacher83c, p. 3)

In folgender Figur ist das Entscheidungsnetz für den Wortexperten "tief" dargestellt. Diese Darstellung geht auf C. Rieger und S. Small (Chuck Rieger, Steve Small: Word Expert Parsing, Univ. Maryland, 1979) zurück.

tief:

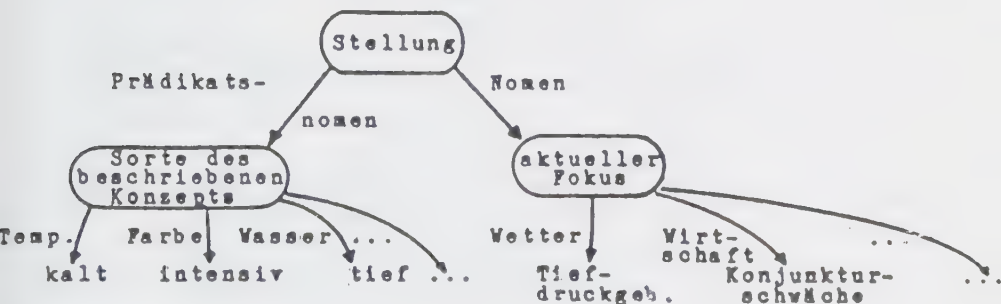


Figure 23. 'Wortexperte (KIT)': (Eimermacher83c, p.3)

Neben Rieger und Small nennt Eimermacher als Referenz für Word Expert Parsing an dieser Stelle:

Kommentiert wird das Netz wie folgt:

"Die Knoten im Entscheidungsnetz stellen Aktionen dar, durch deren Ergebnis der nächste Schritt in Richtung zu den Blättern bestimmt wird. Die Aktionen eines Wortexperten sind Anfragen an benachbarte Wortexperten (WXP's), die alle Bereiche (z.B. Semantik, Syntax und Morphologie) umfassen können".

(Eimermacher83c, p.3) Aus diesem Kommentar geht hervor, dass ein Word Expert Parser

1. ein INTEGRIERTER Parser ist (syntaktisch semantisch morphologisch) und
2. einen DEZENTRALEN Parser darstellt. (Kommunikation zwischen WXP's)

Im Zusammenhang mit "Lexikon" ist vor allem der zweite Punkt von Interesse, denn

"der semantische Kern des Parsers"

(Eimermacher83c, p.11)

ist das WXP-Lexikon.

"Das System benutzt ein Lexikon, das jedem Eingabewort eine semantische Definition zuordnet. Diese besteht aus der

- SRL-Repräsentation für das Wort: Sie ordnet dem Wort i.a. mindestens [**3] einen SRL-Operator zu. [...] Der Operator kann nullstellig sein (z.B. für Namen) oder eine Liste für Argumente haben (wie u.a. Verbergänzungen in der natürlichen Sprache). Die Argumente bestehen zunächst aus Variablen, die mit SRL-Ausdrücken aus dem Kontext aufgefüllt werden müssen. Dazu enthält die Definition ausserdem für jede Ergänzung eine

- Erwartung über den Kontext: Sie gibt an
 - ist die Ergänzung obligatorisch oder optional? Hier sind nur die Ergänzungen gemeint, die vom entsprechenden SRL-Operator gebunden werden. [...].
 - in welchem Bereich soll die Ergänzung gesucht werden? (z.B. next=direkt folgend, any=irgendwo im Satz)
 - eine Sortenbeschränkung für jede Ergänzung.
 - Ausserdem enthält die Kontexterwartung eine Antwortvariable, die mit dem gesuchten Argument belegt werden soll."

(Eimermacher83c, p.11)

Ein Lexikon-Eintrag im WXP-Lexikon stellt also einen Teil des dezentralen Parsers dar und hat folgende Form:

Wortform => SRL-Repräsentation <= Kontexterwartung
--

bzw.

Wortform --> Grundform

(Eimermacher83c, p. 12)

Die beiden Möglichkeiten der Notation eines Lexikon-Eintrages ergeben sich dadurch, dass die 2. Form mit dem einfachen Pfeil lediglich einen Verweis von einer Wortform auf eine Grundform darstellt, z.B. "liegt --> lieg". Von der Implementationsebene her betrachtet, stellt der Pfeil --> einen Infixoperator dar, der das Prädikat repräsentiert, welches lexikalisches "Verweiswissen" enthält.

Der doppelte Pfeil in der 1. Notationsform eines Eintrags hat ebenfalls die Bedeutung eines PROLOG-Infixoperators, welcher lexikalisches Wissen enthält. Eimermacher merkt auf p. 12 (Eimermacher83c) an, dass diese PROLOG-Notation zu folgender Notation in gewöhnlicher Prädikatenlogik äquivalent ist:

"über => LOC: über(*geo) <= next.GEO.*geo."

entspräche der prädikatenlogischen Formulierung:

"lex(über. LOC: über(*geo), next.GEO.*geo)"

(Der nach links gerichtete Pfeil <= markiert die Grenze zwischen den Kontexterwartungen und der SRL-Repräsentation. Der Stern * ist die Kennzeichnung für eine PROLOG-Variable, wobei es sonst in PROLOG üblich ist, Variabeln mit GROSS-Buchstaben zu markieren).

oft

Für den "Satz

'Alfons liegt über Schweden'

(Alfons ist der Name eines Hochdruckgebietes aus einem Wetterbericht)"

(Eimermacher83c, p. 12)

sehen die Lexikoneinträge zu den im Satz vorkommenden Worten wie folgt aus:

```
alfons    => HOCH: alfons
liegt     --> lieg
lieg      => ZUSTAND: liegen(*druck,*loc)
          <= any.DRUCK.*druck 139 any.LOC.*loc
über      => LOC: über(*geo) <= next.GEO.*geo
über      => DLOC: über(*geo) <= next.GEO.*geo
[...]
schweden => LAND: schweden
```

(Eimermacher83c, p. 12)

In Kommentarzeilen macht Eimermacher darauf aufmerksam, dass "über" hier zwei Bedeutungen hat:

- * - Ortsangabe (über Schweden, wo?) *
- * - Ortsveränderung (über Schweden <hinweg>) *

Zusammenfassend kann gesagt werden, dass das KIT-Lexikon eine Sammlung von in PROLOG geschriebenen Kontexterwartungen, SRL-Definitionen und Verweisen darstellt. Diese Definitionen werden als WXP bezeichnet.

Allerdings drückt diese Zusammenfassung nur die Grundidee aus, denn bei der Weiterentwicklung des Konzepts werden (wie bereits erwähnt) die Wortexperten (WXP) nicht mehr direkt im Lexikon eingetragen, sondern aus WSXP (Wortstammexperten) und Endungen generiert. (Eimermacher83b, p.1ff.)

Mit andern Worten formuliert bedeutet dies, dass das ursprüngliche WXP-Lexikon ein Vollformenlexikon war und durch ein Stammformenlexikon ersetzt wurde.

"Der semantische Parser aus (Eimermacher83c) verwendete - wie der ursprüngliche WXP-Ansatz aus Rieger/Small [siehe oben] - ein Vollformenlexikon. Mit Hilfe von Metaeinträgen konnte die Redundanz im Lexikon verringert werden, aber es fehlte die morphosyntaktische Analyse und damit der Zugriff von verschiedenen Wortformen auf die Informationen über einen gemeinsamen Wortstamm" (Eimermacher83a, p.10)

Durch die Einführung eines Stammformenlexikons wurde die Redundanz in den Einträgen reduziert, aber der Regelapparat zur Bildung und Erkennung von Vollformen musste entsprechend erweitert werden. Für die Lexikoneinträge bedeutet dies, dass

"Die auf das Wortstammwissen reduzierten Lexikoneinträge" (Eimermacher83b, p.7)

zu Wortstammexperten (WSXP) werden und nicht mehr vollständige Wortexperten (WXP) sind.

Der Unterschied des weiterentwickelten Konzepts zum ursprünglichen besteht auch darin, dass die dezentralen Parserkomponenten (welche mit den Lexikoneinträgen identisch sind) nicht mehr aus Wortexperten (WXP), sondern aus Wortstammexperten (WSXP) bestehen.

Dabei ist es wichtig zu wissen, was unter "Stammform", "Wortstamm" und "Grundform" verstanden wird. Eimermacher unterscheidet diese 3 Begriffe deutlich:

"- Stammform: Die von den Endungen befreite Form

(1.Sg.	2.Sg.	Präteritum	Konj.Prät.	Part.Perf.Pass.):
lauf	läuf	lief	lief	(ge)lauf(en)
lieg	lieg	lag	låg	(ge)leg(en)

- Der Wortstamm soll in diesem Bericht nicht im linguistisch-historischen Sinn verstanden werden - es gibt hier keinen gemeinsamen Wortstamm für 'liegen' und 'Lage'. Wortstamm werden die Stammformen genannt, für die (etwa wegen einer starken Flexion) ein eigener Lexikoneintrag verwendet wird.

- Grundform: für jede Wortart willkürlich ausgewählter Wortstamm, z.B. bei Verben: Präsensstamm.

Die folgende Tabelle zeigt bei der schwachen Flexion einen gemeinsamen Wortstamm für mehrere Wortarten (Strömung, strömt) und für die starke Flexion mehrere Wortstämme innerhalb einer Wortart (lieg, lag) mit einer gemeinsamen Grundform (lieg): "
(Eimermacher83b,p.5)

WORTFORM	WORTART	STAMMFORM	WORTSTAMM	GRUNDFORM
strömt	Verb	ström	ström	ström
strömende	Partizip	ström(end)	ström	ström
strömungen	Nomen	strömung	ström	ström
lieg	Verb	lieg	lieg	lieg
lagen	Verb	lag	lag	lieg
lagen	Nomen	lage	lage	lage

(Eimermacher83b, p.5)

Mit Hilfe dieser Unterscheidungen macht Eimermacher nochmals deutlich, wie in KIT die Lexikoneinträge konzipiert sind und wie sie durch die Deflexionsanalyse erhalten werden:

"Im Stammformenlexikon werden nicht mehr die Wortformen selbst gespeichert. Sie werden durch die Deflexionsanalyse auf eine Stammform zurückgeführt (lagst => lag). Für jede Stammform kann das Lexikon einen Verweis auf die Grundform enthalten (lag --> lieg). Bei schwacher (regelmässiger) Flexion können auch die Stammformeinträge entfallen, und das Wort ('sagte') wird in der Deflexionsanalyse unmittelbar auf die Grundform als einzigen Wortstamm (sag) zurückgeführt. Dies gilt auch für die regelmässige Suffixbildung: 'Strömung' hat als Wortstamm den Verbstamm 'ström'. Bei starker (unregelmässiger) Flexion werden die Wortformen nur auf die Stammform zurückgeführt ('lagst' => lag). D.h. 'lag' wird als eigener Wortstamm eingetragen. Für ihn enthält das Lexikon dann den Verweis auf die Grundform (lag --> lieg), i.a. mit zusätzlichen Angaben über den Wortstamm, z.B. lag --> lieg;Präteritum. Das sprachliche Wissen über die Worte ist somit in beiden Fällen auf Wortstammeinträge konzentriert"
(Eimermacher83b, p.6/7)

Wie Wortstammexperten aussehen können, zeigt die folgende Abbildung.

Wort- stamm	(SORTE,	Argumente	Syntaktische Merkmale<2>)
zieh	(EREIGN,	(Agens.MO).(DLoc.DLOC).nil,	V;prä;intrans;*.*.zog.zog)
lieg	(ZUSTAND,(Agens.MO).(Loc.LOC).nil,		V;prä;intrans;*.*.lag.leg)
hoch	(HOCH , nil,		Sort;2.nil).
hoch	(ZUSTAND, (*.MO).nil,		Adj;pos;*.hoeh.hoech).

Beispiele von Wortstammexperten (WSXPs)

Figure 24. 'Wortstammexperte (KIT)': (Eimermacher83b,p.6)

Diese Figur macht deutlich, welche Kategorien im KIT-Lexikon der Wortstämme abgespeichert sind, die sich mit Kategorien vergleichen lassen, wie sie in den anderen untersuchten Systemen z.T. auch vorkommen. Es gehören dazu:

- "die Sorte
 - mögliche Ergänzungen, beschrieben durch deren zulässigen Sortenbereich und Tiefenkasus (TK): MO und LOC sind sozusagen die Slots des Konzepts 'lieg' auf sprachlicher Ebene."
- (Eimermacher83b,p.6)

Auch die syntaktischen Merkmale im Eintrag gehören zu diesen Kategorien. Die in der Figur von M. Eimermacher dazu gesetzte Anmerkung <2> lautet:

"Das erste syntaktische Merkmal ist die Wortart (V='eigenständige Verbform' - d.h. im Deutschen: Aktivform des Präsens oder Imperfekt vs. Infinitiv- und Partizipialbildungen z.B. für das Passiv, Sort=Sortenbezeichner - eine Untergruppe der Nomen). Das letzte Merkmal ist die Flexionsart: Für starke Verben werden die Stammformen angegeben, für schwache nur die freie Variable (*), die durch die Standardkonjugationsformen instantiiert wird."

(Eimermacher83b,p.6)

*

-EVAR:

Die in der Synopse unter EVAR genannten Kategorien sind im Lexikon nicht als lineares Abfolgescema angeordnet, sondern bilden eine syntaktisch organisierte Baumstruktur. Die gesamte Syntax dieser Struktur wurde in (1.1.1(f)) abgebildet. Figur 3 zeigt für den semantischen Teil des Eintrags dessen hierarchische Struktur. Für die Struktur des semantischen Teils des Eintrags ist auch Figur 11 relevant.

Da auf das Lexikon im System EVAR in (2.3.2) nochmals eingegangen wird, mögen diese Hinweise vorläufig genügen.

*

-SADAW:

Die in der Synopse genannte Kategorienfolge

- "Identifikationswortlaut",
- "Bedeutungsnummer" und
- "Informationsteil"

ist eine grobe Einteilung der im Wörterbuch SADAW gespeicherten Information. Eggers et al. bemerken zu diesen Kategorien:

"Bei der Bedeutungsnummer handelt es sich um eine Zahl, die zur Unterscheidung von Lexikoneinträgen mit gleichem Identifikationswortlaut dient.

Beispiel:

IDENTIFIKATIONSWORTLAUT	BEDEUTUNGSNUMMER	ERKLÄRUNG
LAUT	1	Singularstamm des Subst.
LAUT	2	Stamm des Verbs 'lauten'

Durch Identifikationswortlaut und Bedeutungsnummer ist ein Lexikoneintrag eindeutig identifiziert.

Im Informationsteil sind die morphologischen, syntaktischen und semantischen Informationen zum Identifikationswortlaut untergebracht, sowie dessen Grundform."
(Eggers80, p.56/57

In Bezug auf den Informationsteil muss einschränkend gesagt werden, dass in SADAW hauptsächlich morphologische und syntaktische Information gespeichert ist, weniger semantische. Als Beispiel möge ein in (Hess83,p93) dargestellter Verbstammeintrag dienen:

Verbstammeintrag (erste Verbkarte)

Spalte	Inhalt
1-19	Identifikationswortlaut
20	Angaben zur Wortlautfolgekarte
21	Markierung f. lexikal. Mengenbeschränkung
22	Fachgebietmarkierung
23	frei
24	Bedeutungsnummer
27	Länge des trennbaren Präfixes
28-52	Angaben zur Flexion wie Endungen des Präsens, Bildung des Präteritums, Passivbildung, Stammtyp; Angaben syntaktischer Art wie Anschluß von Daß- oder Fragesatz, Wortklassen-übergang
53-55	Präpositionale Valenzen
64-65	obligatorische Angabe bei Infinitivanschluß
66	Angabe der Valenz, die der Inf.Satz befriedigt

Verbstammeintrag (zweite Verbkarte)

Spalte	Inhalt
62-63	Allgemeine Angaben zur Bindung von Verbzusätzen
64-65	Verbsimplexnummer
66-68	Klasse von Verbzusätzen, die das Verb haben kann

Figure 25. 'Verbstammeintrag (SADAW)': nach (Hess83,p.93)

Aus Platzgründen ist es nicht möglich, die in der Benutzeranleitung für SADAW im Detail dargestellten Kategorien aufzulisten. Es sei aber auf (Eggers80) hingewiesen, wo sich ebenfalls genaue Angaben finden lassen.

*

-ALEXSYS:

Die in der Synopse genannten 7 Informationsklassen

- Worteintrag (WE)
- Phonetik/Phonologie
- Wortklasse
- Morphologie
- Syntaktische Angaben
- Semantische Angaben
- Pragmatischer Kommentar

sind in Figur 15 erläutert. Eine ausführlichere Liste der Kategorien zu den 7 Informationsklassen ist in (Hess83), p.198-201, zu finden. (vergl. auch p.219).

Die Struktur der morphosyntaktischen Angaben (MSA) wurde schon in (1.1.2(c)) im Zusammenhang mit der Implementation variabler Abfolgescemata beschrieben. (vergl. Figur 2).

Einen Überblick über die wichtigsten Kategorien geben Hess et al. auf Seite 219. Die dort dargestellten Kategorien sind:

```

WG  Wortlaut
L   Lemma
P   Phonologie
WK  Wortklasse
F   Flexionsmorphologische Angaben
    Kasus
    Numerus
    Genus
    Steigerungsstufe
    Person
    Numerus
    Modus
    Tempus
    Aktiv/Passiv
    Perfektbildung
D   Derivationsangaben
OS  Syntaktische Umgebung
TS  Tiefenkasus
Semantische Angaben
    DEF      (Semantische Definition)
    SEMKON   (Semantisch relevanter Kontext)
    ANW      (Anwendungsbeispiele)
    SEREL    (Semantische Relationen)
    PRIM     (Semantische Primitive)
    FG       (Fachgebietsmarkierung)

```

(Hess83, p.219)

*

-TEAM:

Die in der Synopse genannten Informationsteile

- "Eintragskopf" und die 8
- Sprachen

sind nicht Kategorien, sondern Blöcke von Kategorien. Die einzelnen Kategorien sind in den Figuren 18 und 19 dargestellt.

Als Beilage zum Fragebogen wurden auch die Details der Felderbelegungen der Kategorien 00 bis 19 angegeben.

Diese Details bestehen für den Eintragskopf (00-09) aus folgenden Angaben:

- 00 Eintragsadresse. (OBLIGAT)
2 Grossbuchstaben (keine Umlaute), 4 Ziffern.
Beispiel: ZZ0011
- 01 (frei)
- 02 (frei)
- 03 Qualitätskennung des (mehrsprachigen) Eintrags,
d.h. Bewertung der Äquivalenzbeziehungen;
bei einsprachigen Einträgen Angabe der Einsprachigkeit
1 Buchstabe.
Beispiele:
E = Einsprachiger Eintrag
A = Äquivalenz nicht gesichert bzw. Äquivalenzgrad nicht
definiert; "Arbeitsbenennung"
d = Äquivalenz gesichert bzw. Äquivalenzgrad definiert;
endgültig geklärte Benennung
- 04 Zeitpunkt der Erarbeitung oder Eingabe des Eintrags, (OBLIGAT)
oder Zeitpunkt einer gezielten Überprüfung, Aktualisierung
oder Erweiterung.
4 Ziffern
Beispiel: 0883 = August 1983
- 05 Teilbestandskennung, (OBLIGAT)
d.h. Kennung des Teilbestandes, dem der Eintrag angehört
oder zugeführt werden soll.
4 Grossbuchstaben.
- 06 Fachgebietskennung(en).
1 Grossbuchstabe (kein Umlaut), 4 Ziffern.
Beispiel: E0115, E0323
- 07 (frei)
- 08 (frei)
- 09 Geltungsbereichskennung(en).
Möglichst kurz fassen!
Beispiel: EWSA; EWSD

[e]

16. FRAGE:

Besteht die Möglichkeit, Einträge zu verknüpfen? (Verknüpfung
Lemma-Formen/ Thesaurusrelationen/ Frames/ andere

Diese Frage steht in Zusammenhang mit den in (1.3.1) beschriebenen
"analysierbaren Teilen des Lexikons", d.h. der Mikro-, der Makro-
und der Verflechtungsstruktur (1.3.1(c)).

ANTWORTEN:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
LEMMA-FORMEN KASUS- RAHMEN		LEMMA-FORMEN	NEIN	NUR DURCH EXPLIZI- TEN ZUGRIFF	LOGISCHE VER- KNÜPFUNG MIT SELEKTIONS KRITERIEN

Das Wortstamm-Lexikon von KIT basiert prinzipiell auf Eintragsverknüpfung, weil die Lexikoneinträge in Form von kommunizierenden Wortstammexperten (WSXP) Teile des dezentralisierten Parsers sind.

Die unter HAM genannte Verweismöglichkeit auf Kasusrahmen zeichnet sich dadurch aus, dass dieser Verweis vom Wortlexikon in eine andere lexikalische Wissensquelle, das Kasusrahmenlexikon, zeigt.

Die bei TEAM genannte Möglichkeit logischer Verknüpfung mit Selektionskriterien bezieht sich auf die Möglichkeit, über ein Suchprofil eine Menge von Einträgen nach bestimmten Kriterien zu selektieren.

[f]

20. FRAGE:

Ist es Ihnen möglich, einen typischen Auszug aus dem Wörterbuch und eine Erklärung des Codes beizulegen?

ANTWORTEN:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
-	-	-	-	separat	Kateg. schema

Die Antwort auf diese Frage wurde in allen 6 Fällen in Form von Begleitmaterial zum retournierten Fragebogen erteilt.

*

Ein zusammenfassender Vergleich der linguistischen Konzeption der 6 Lexikonsysteme ist deshalb schwierig, weil "linguistische Konzeption" ein doch eher abstrakter Begriff ist. Es ist deshalb sinnvoller, keinen pauschalen Vergleich durchzuführen, sondern jeweils nur in Bezug auf nennbare Kriterien, wie sie in den eben behandelten 6 Fragen [a] - [f] ausgedrückt wurden, zu vergleichen. In Bezug auf diese 6 Kriterien

- Zweck [a]
- morphologischer Typ [b]
- Wortklassenschema [c]
- Eintragskategorien [d]

- Verknüpfungen [e]
- Codes [f]

bieten die 6 Synopsen mit ihren Kommentaren einen Vergleichsraster, welcher für spezifische Vergleiche auf Teilgebieten wiederverwendet werden kann. Für den in (2.2.3) gemachten Vergleich semantischer Eintragskategorien werden die Antworten und Kommentare weiterverwendet.

Der Vergleich ist damit nicht Selbstzweck, sondern Basis für ausgewählte spezifische Vergleiche. Weil ich mich in (2.2.3) dafür entschieden habe, semantische Kategorien zu vergleichen, wird dort u.a. auf die Antworten und die Kommentare zur Frage Nr.4 (Informationsfelder pro Eintrag) zurückgegriffen, um aus den Eintragskategorien die semantischen Kategorien zu extrahieren. Damit ist der Sinn des synoptischen Vergleichs evident: Die Synopsen sollen einen problemlosen Zugriff auf charakteristische spezifische Merkmale der 6 untersuchten Systeme ermöglichen.

*

3.1.2 Datengrundlage.

[a]

9. FRAGE:

Sind ausser der codierten Lexikoninformation noch andere Informationen codiert? (Affixlisten/ andere)

Die Antworten auf die 2 Fragen in dieser Gruppe sollen zwei verschiedene Dinge zeigen:

- a) Die Homogenität bzw. Heterogenität der lexikalischen Wissensquellen in den betreffenden Systemen (Frage Nr. 9) und
- b) Die Primärquellen des verwendeten lexikalischen Wissens (Frage Nr. 19)

ANTWORT:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
Affix- listen	-	-	Präfixe Suffixe Fugen	-	-

Auf Zusatzinformation ist auch das Lexikon von KIT angewiesen, weil es sich um ein Stammformenlexikon handelt. Diese Zusatzinformation liefern die:

- Syntaktische Kongruenztabelle (vergl. Anmerkung [**2] zu Teil 3) und die

- Endungstabelle für die Deflexion.

Auch Wortartenverband und Sortenverbände können dazugezählt werden.

*

Für das Lexikon in HAM-ANS sind die zusätzlichen lexikalischen Wissensquellen bereits erwähnt worden.

*

Die vollständige Antwort für SADAW lautet:

"Präfixe, Suffixe, Fugen sind ebenfalls Einträge"

[b]

19. FRAGE:

Welche konventionellen Wörterbücher dienen als Grundlage ? (für den Wortschatz/ für die Codierung/ andere Informationen)

Die Frage weist einen kleinen Schönheitsfehler auf, weil nur nach konventionellen Wörterbüchern gefragt wurde, aber in den Antworten werden auch maschinenlesbare Primärquellen genannt.

ANTWORTEN:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
-	-	Grundwortschatz, Helbig/Schenkel, Schumacher, Duden	Wahrig, Grundwortschatz, eigene Korpora	CONDOR LIMAS Mackensen MOLEX SADAW SDW VALEX Wahrig	-

Die fehlenden Angaben bei HAM und bei KIT lassen sich teilweise dadurch erklären, dass diese beiden Lexika umfangsmässig nicht sehr gross sind und sich deshalb das Problem grosser Primärdatenmassen nicht stellt. Im Falle von TEAM ist es so, dass die Daten von den Terminologen und Übersetzern des "Sprachendienstes" der Firma laufend in grossen Massen produziert und aktualisiert werden.

*

Bei dem unter EVAR angegebenen "Grundwortschatz" handelt es sich teilweise um eine Übertragung eines Dialog-Minimalwortschatzes aus dem Englischen (M.J. Kelly, A. Chapanis, Limited vocabulary natural language dialog. In: International Journal of Man-Machine Studies, Vol. 9, No. 4 (June77), p.479ff) In (Müller-Meernach81), p.11ff. finden sich nähere Angaben zur Grundwortschatzproblematik.

Auf die in EVAR verwendeten Valenzwörterbücher wird in (2.2.3) und in (2.3.2) noch eingegangen.

*

Zur Primärdatengrundlage des Wörterbuchs SADAW kann folgendes gesagt werden:

Das "maschinelle Wörterbuch baute ursprünglich auf einem lexikalischen Inventar auf, das seitens der Stichwörter weitgehend dem "einbändigen" Lexikon von G. Wahrig in einer Ausgabe zu Ende der 60-er Jahre entsprach."
(Zimmermann83, p. 7)

"Inzwischen ist dieses maschinelle Wörterbuch auf rd. 150 000 derartige "Stämme" [**4] angewachsen, wobei Erweiterungen vorwiegend aus den Fachgebieten der behandelten Texte (Recht, Patentwesen, Sozialwissenschaften) stammen."
(Zimmermann83, p. 7)

*

Die unter ALEXSYS genannten Wörterbücher entsprechen den in das kumulierte Lexikon aufgenommenen Einzelllexika. (Vergl. die Tabelle in Figur 14)

*

3.1.3 Implementation.

Die in dieser Gruppe zusammengefassten Fragen betreffen zu einem grossen Teil die Gegenüberstellung "File-System" versus "Datenbank-System". (Vergl. dazu Anmerkung [**1] zu Beginn von Teil 3). In zwei Fällen (EVAR und ALEXSYS) wurde in den Antworten explizit auf eine bevorstehende Umstellung von Filetechnik auf Datenbanktechnik hingewiesen.

Ein im Hinblick auf "Implementation" interessanter Aspekt wird von M. Eimermacher im Antwortbrief auf die Fragen zum Lexikon KIT festgehalten:

"Zum modifizierten Word-Expert-Parser muss ich auf IAB9 [**5] verweisen, da der Fragebogen sich zum grossen Teil auf Implementationsdetails bezieht, von denen wir in PROLOG abstrahieren können. Z.B. interpretiert PROLOG eine Menge von "Fakten" auch als Datenbank, obwohl sie als gewöhnliche Datei oder Modul eingegeben werden kann."
(Eimermacher85)

Leider ist es nicht immer möglich, von der Art der Implementierung zu abstrahieren. Diese Abstraktionsmöglichkeit hängt z.T. von der Projektphase ab. In der Phase des Prototyping kann durchaus von Implementierungsdetails abstrahiert werden, aber in späteren Phasen, wenn die Datenmenge grösser wird und der Prozess optimiert werden soll, kann auch in PROLOG von der Art der Implementation

nicht mehr als "quantité négligeable" gesprochen werden.

Um die Antworten auf die folgenden Fragen [a] bis [e] unter dem hier interessierenden Aspekt der Verwaltung lexikalischer Daten zu verstehen, sei an dieser Stelle nochmals auf Anmerkung [**1], Teil 3 hingewiesen. Die aus den Fragen nicht ersichtliche Systematik wird dort illustriert.

[a]

6. FRAGE:

Ist Ihr Verfahren fileorientiert oder datenbankorientiert ?

ANTWORTEN:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
DB	[EXP]	FILE --> DB	FILE	FILE --> DB	FILE und DB

Das unter KIT ergänzte EXP soll darauf hindeuten, dass es sich bei diesem Lexikon um WortEXPerTen handelt, welche als PROLOG-Wissensbasis implementiert sind.

DB bedeutet "Datenbank", wobei die verwendeten Systeme folgende sind:

- HAM: Eine in UCI-LISP eingebettete Eigenentwicklung mit Worteinträgen als FEATURE-VALUE-PAAREN und einem hash-Zugriffsalgorithmus.

- EVAR: HBASE. Das System geht nach Angabe von (Endres86,p.95) auf H.G. Barrow vom A. I. Center des SRI (Stanford Research Institute) zurück und wurde 1984 durch C. Beckstein von der Universität Erlangen/Nürnberg auf Interlisp angepasst, später

"durch Mitarbeiter des Lehrstuhls [**6] für Franz-Lisp angepasst, wo es auf einer VAX 11/730 und VAX 11/750 unter ULTRIX für diese Arbeit [**7] zur Verfügung stand." (Endres86,p.7)

- ALEXSYS: SESAM, ein DBMS der Firma Siemens.

- TEAM: TERMAS, ein Programmpaket der Firma Siemens (Terminologie Abfrage System); ein komplementäres Programm, TERMSEL, dient der Selektion grösserer Term-Mengen u.a. zur Wörterbuchherstellung.

[b]

7. FRAGE:

- Aus welchen Feldern ist ein logischer Record aufgebaut ?

Teilantwort:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
[x]	[x]	Textfile	-	-	Kate- gorien- schema

Die unter HAM und KIT eingesetzten "x" bedeuten, dass die Frage für diese Systeme nicht relevant ist.

Da bei SADAW die Recordarten den Wortarten entsprechen (vergl. übernächste Teilantwort), muss auf die bereits erwähnten Detailangaben in (Eggers80) und in der beigelegten Benutzeranweisung verwiesen werden.

Der Recordaufbau in ALEXSYS wurde schon dargestellt.

- Welche Recordarten benutzen Sie

Teilantwort:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
[x]	[x]	[x]	-	4 ISAM- FILES für Wortart- gruppen: Nomen, Verben, Qualifier, Fkt. wörter	-

Hier gilt das Gleiche wie für die vorige Frage.

- Entsprechen Ihre Recordarten den Wortarten?

Teilantwort:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
[x]	[x]	[x]	ja	nein	nein

*

[c]

8. FRAGE:

Falls datenbankorientiert:

- welchen DB-typ verwenden Sie? (CODASYL/ Baumhierarchie/
Inverted File/ Relational/ andere)

ANTWORTEN:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
Eigen- entwklg in LISP einge- bettet	Prolog- basis	neu: HBASE adaptiert	[x]	neu: SESAM	TERMAS

[d]

11. FRAGE:

Mit welchen Zugriffsalgorithmen operieren Sie? (VSAM/ hash/ B-Baum/
andere)

ANTWORTEN:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
hash	[PROLOG]	noch unklar	B-Baum	ISAM	VSAM

[e]

12. FRAGE:

Welche Felder können als Zugriffsschlüssel verwendet werden? (nur
ein Schlüssel/ alle Felder/ andere)

ANTWORTEN:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
nur ein Schlüssel	[PROLOG]	nur ein Schlüssel (Wortnr.)	nur ein Schlüssel	Direkt- zugriff: Wortform	Abfrage: alphabe- tisch in ver- langter Sprache od. nach Eintrags- nummer

[f]

10. FRAGE:

In welcher Programmiersprache operieren Sie über dem Wörterbuch?

ANTWORTEN:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
UCI-LISP	[PROLOG]	SIMULA neu: LISP	FORTTRAN Ling. Untprg	PL1	ASSEMB LER

Die Kommentare zu den Fragen, welche die Implementation betreffen, fielen kürzer aus als diejenigen zur Fragengruppe "Linguistische Konzeption", weil der Schwerpunkt vorliegender Arbeit linguistisch und nicht informatisch ausgerichtet ist.

3.1.4 Quantitative Merkmale und Restriktionen.

Beim Grössenvergleich von maschinenlesbaren Wörterbüchern ist es wichtig festzuhalten, in welcher Einheit gemessen werden soll. Wird von Stammformen oder von Grundformen ausgegangen, so können sich die Zahlen zwar bescheiden ausnehmen, aber wenn man diese Angaben auf Vollformen umrechnet, kann man (als Faustregel) etwa die zehnfache Zahl von Vollformen ansetzen.

In den Antworten zur 4. Fragengruppe sind neben Grössenmerkmalen auch quantitative Restriktionen dokumentiert: Wortlängen- und Eintragslängenrestriktionen.

Bei den Antworten zur Frage Nr. 15 beziehen sich diese teilweise auf eine Längenrestriktion, teilweise auf die effektive maximale Belegung, wobei aber im Einzelfall sofort klar ist, was gemeint war.

[a]

13. FRAGE:

Wie gross ist ihr Wörterbuch? (Anzahl Einträge/ Anzahl Grundformen/ Anzahl abgeleitete Formen/ andere Grössenmerkmale)

ANTWORTEN:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
ca 1500 GF	einige 100 SF	ca 5000 VF	ca 110'000 GF	ca 300'000 VF/GF	ca 1,5-2 Mio VF

(SF=Stammformen, GF=Grundformen, VF=Vollformen)

(b)

14. FRAGE:

Maximale Wortlänge ?

ANTWORTEN:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
nicht begrenzt	[PROLOG]	nicht begrenzt	36 Buchstaben	40 Zeichen	4000 Byte gilt für Maxlänge des ganzen terminolog. Eintrags

[c]

15. FRAGE: Maximale Anzahl von Informationskategorien (Codes) pro Lexem?

ANTWORT:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
ca 7	[PROLOG]	beliebig	256 Byte plus 2 Wort- laute	wort- klassen- abhängig, für Verben: ca 40	98

3.1.5 Umfeld.

Die Fragen zum Thema "Umfeld" sind vor allem dort relevant, wo das Lexikon für verschiedene Zwecke eingesetzt und von verschiedenen Benutzerkreisen benutzt werden könnte.

[a]

17. FRAGE:

Welche LDV-Programme (oder Systeme) benutzen Ihre Wörterbücher?

ANTWORTEN:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
HAM-ANS Ober- flächen- transfor- mationen auf Symbolics Rechnern		-	eigenes System innerhalb von SUSY	z.Zt. keine	Progr. system TEAM

Für KIT-Anwendungen vergl. p.76 (KIT-NATAN).

[b]

18. FRAGE:

Existiert eine Schnittstelle zur DIN 2341 (z.B. MATER)?

ANTWORTEN:

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
nein	nein	-	-	z.Zt. nicht	in Arbeit

Die vielen fehlenden und auch die negativen Antworten zeigen, dass gerade auf diesem Gebiet ein grosses Entwicklungspotential vorhanden ist.

Als Entwicklungsperspektive kann auf das Konzept "Wortdatenbank" hingewiesen werden. Vergl. dazu (1.1.2.(d)). Eine Datenbank soll sich ja gerade durch ihre Datenunabhängigkeit auszeichnen, was heissen soll, dass verschiedene Applikationen dieselbe Datenbasis zu unterschiedlichen Zwecken benutzen können.

In der Praxis fehlen aber -wie die Antworten zeigen- die Schnittstellen.

3.1.6 Hardware.

5. FRAGE:

Welchen Rechner benutzen Sie?

ANTWORTEN:

HAM	KIT	EVAR	SFB100	ALEXSYS	TEAM
DEC-10	IBM 4381	VAX11-730 CYBER 895	Siemens 7760	IBM 370/ 168	Siemens 7.5xx (BS2000)

*

3.2 * VERGLEICH SEMANTISCHER KATEGORIEN.

Bei diesem Vergleich handelt es sich um eine Spezialisierung des in (3.1) durchgeführten synoptischen Vergleichs der 6 untersuchten Lexikonsysteme. Aus den in (2.1.*) genannten Gründen wurden für diesen spezifischen Vergleich die semantischen Kategorien gewählt.

Der Vergleich von semantischen Kategorien aus Systemen mit unterschiedlichem Hintergrund bietet den Vorteil, "Semantik" aus unterschiedlichen Perspektiven betrachten zu können. Dadurch ergibt sich die Möglichkeit, die zu Ende des Kapitels (2.1.*) erwogene Relativierung von Semantikmodellen konkreter zu fassen. Damit sich aber diese vorteilhafte Möglichkeit nicht in einen Nachteil verwandelt, muss auf jeder Stufe des Vergleichs berücksichtigt werden, dass "Semantik" in den 6 Systemen durchaus unterschiedlich verstanden wird. Der Vergleichsraster ist deshalb so beschaffen, dass gerade auch der semantische "Background" berücksichtigt wird.

Der Vergleich semantischer Kategorien wird deshalb auf verschiedenen Ebenen durchgeführt:

- a) Identifizierung der semantischen Kategorien in den einzelnen Systemen (3.2.1).
- b) Einbezug des semantischen "Backgrounds" (semantisches Bezugssystem) (3.2.2).
- c) Vergleich der aus Kategorien und Bezugssystem ableitbaren semantischen Konzeptionen (3.2.3).

Dieses gestufte Vorgehen unter Einbezug des semantischen Hintergrunds verhindert, dass Kategorien verglichen werden, die, isoliert betrachtet, nicht viel miteinander zu tun haben.

3.2.1 Identifizierung der semantischen Kategorien

Die semantischen Kategorien können aus dem im vorigen Kapitel aufgezählten Total der Eintragskategorien entnommen werden (vergl. Synopse und Kommentar zur 4. Frage des Bogens). Im folgenden werden die semantischen Kategorien nach Projekten geordnet aufgelistet:

HAM:

Von den zum Zeitpunkt der Untersuchung im Wortlexikon angegebenen 8 Eigenschaften (vergl. Figur 8) können

- Synonym und
- Kasusrahmen

als semantische Kategorien bezeichnet werden. Von diesen beiden Kategorien ist "Kasusrahmen" die komplexere Einheit und soll deshalb näher erläutert werden. Diese Kategorie verknüpft das Wortlexikon mit dem Kasusrahmenlexikon, indem sie im Wortlexikon lediglich als Verweis auf einen Eintrag im Kasusrahmenlexikon auftritt, was Hoepfner wie folgt beschreibt:

"Die Wissensquelle Kasusrahmen-Lexikon in HAM-ANS enthält alle für Verben vorgesehenen Kasusrahmen. Im Lexikoneintrag einer Verbgrundform, deren Semantik durch Kasusrelationen beschrieben wird, ist ein Verweis auf den zutreffenden Kasusrahmen enthalten. Dadurch ist die Verwendung eines bestimmten Kasusrahmens für mehrere Verben während der Analyse möglich, ohne dass damit eine Gleichbehandlung in der Auswertung ganzer Äußerungen präjudiziert wird. Ein Beispiel dafür sind die beiden Verben 'beschleunigen' und 'bremsen', für die bei der Analyse derselbe Kasusrahmen verwendet wird, der einen obligaten AGENT vom Typ 'Fahrzeug' und einen fakultativen LOCATIVE vom Typ 'Verkehrsweg' verlangt. Die augenblicklichen Einträge des Kasusrahmen-Lexikons wurden für die bisher verarbeitbaren Verben aufgebaut und stellen kein abgeschlossenes System dar".
(Hoepfner82,p.217)

Die in HAM-ANS verwendeten Kasus werden in (Hoepfner82) genannt und waren damals folgende:

AGENT
OBJECTIVE
SOURCE
GOAL
TIME
PATH
INSTRUMENT

"Diese Liste erhebt keinen Anspruch auf Vollständigkeit - beispielsweise fehlt noch ein Recipient- oder Benefactive-Kasus - sondern ist für die bisher durch Kasusrahmen beschriebenen Verben aufgestellt worden."
(Hoepfner82,p217)

Die Kasus sind in die den Verben zugeordneten Rahmen integriert und haben je folgende drei Eigenschaften:

- "1. den Kasusnamen (als syntaktische Markierung in SURF, z.B. agent:, goal:)"
- "2. eine Kennzeichnung, ob dieser Kasus obligat oder fakultativ ist (O bzw. F)" und

"3. semantische Restriktionen, die eine syntaktische Teilstruktur erfüllen muss, um die Kasusrolle übernehmen zu können."

(Hoeppner82,p.217)

Durch die erste und die dritte der genannten Eigenschaften ist angedeutet, dass die Kasusrahmen eine Brückenfunktion zwischen der semantischen Repräsentationssprache SURF und der syntaktischen Satzstruktur erfüllen. In diesem Zusammenhang sei an die von C. Schwind erwähnten Übersetzungsregeln aus "Semantikkonzepte in der Künstlichen Intelligenz" erinnert (vergl. Kapitel 2.1.*). Bei Hoeppner ist das explizit formuliert:

"Kasusrahmen sind in der semantischen Repräsentationssprache SURF formuliert, in der in HAM-ANS auch vollständige Äusserungen repräsentiert sind. Dies hat den Vorteil, dass Interpretationsverfahren für die semantischen Restriktionen weitgehend durch Inferenzen aus der Auswertungskomponente gegeben sind. Die Bedingungen, dass beispielsweise die NP 'ein gelber Personenwagen' ein Objekt der Klasse der Fahrzeuge beschreibt, und damit ein Kandidat für die AGENT-Rolle des Verbs 'einbiegen' ist [...], kann wegen der Transitivität der Oberbegriffsrelation im semantischen Netz leicht nachgewiesen werden."
(Hoeppner82,p.218)

In der folgenden Figur ist der in der Sprache SURF als SURF-Ausdruck notierte Kasusrahmen für Verben des Typs 'einbiegen' dargestellt.

```

(rl-s:
  agent:    (d-l: role-marker: 0
              restrictions:
                (d-o: OR
                  (lambda: x1 (af-a: ISA x1 FAHRZEUG))
                  (lambda: x1 (af-a: ISA x1 MENSCH))))
  locative: (d-l: role-marker: F
              restrictions:
                (lambda: x1 (af-a: ISA x1 VERKEHRSWEG)))
  source:    (d-l: role-marker: F
              restrictions:
                (lambda: x1 (af-a: ISA x1 VERKEHRSWEG)))
  goal:      (d-l: role marker: F
              restrictions:
                (lambda: x1 (af-a: ISA x1 VERKEHRSWEG)))

```

Kasusrahmen für Verben vom Typ *einbiegen*

Figure 26. 'Kasusrahmen (HAM-ANS)': (Hoepfner82, p.218)

Bei dieser Kasusrahmen-Notation ist aber zu beachten, dass sie nicht der im Lexikon notierten entspricht:

"Im Kasusrahmen-Lexikon stehen die einzelnen Kasusrahmen allerdings nicht in dieser Form, sondern als Anweisungen zur Konstruktion eines Kasusrahmens auf der Grundlage der aktuellen Syntaxdefinition von SURF. Dadurch ist gewährleistet, dass sich alle Modifikationen in der Sprache SURF in den verwendeten Kasusrahmen niederschlagen." (Hoepfner82, p.218)

Die Generierung der Kasusrahmen im Laufe der Analyse wird mit Hilfe einer SEMANTISCHEN ATN-Grammatik durchgeführt (ATN-Grammatiken wurden von Woods ursprünglich als syntaktisches Analyseinstrument entworfen). Das Verfahren ist in Hoepfner82 beschrieben.

*

Im System HAM-ANS sind abgesehen von der wichtigen Kategorie "Kasusrahmen" weitere semantische Kategorien im Gebrauch, die aber nicht in der im System umfangreichsten lexikalischen Wissensquelle

Wortlexikon verzeichnet sind. Dazu gehören

- Paraphrasen von Redewendungen (Syntagmenlexikon) und
- semantische Netzrelationen

Auf diese und weitere semantische Kategorien (Scripts, Inferenzregeln) wird aber im Zusammenhang mit dem Systembackground (semantisches Bezugssystem) noch eingegangen, wobei die bereits erwähnte Heterogenität der lexikalischen Wissensquellen in HAM-ANS (2.2.2.1) zu berücksichtigen ist:

- Wortlexikon,
- Syntagmenlexikon,
- Kasusrahmenlexikon,
- Flexionssuffixe,
- Semantische Netze,
- Semantische Repräsentationssprachen SURF und DEEP

*

KIT:

Dieses Lexikon stellt insofern einen Spezialfall dar, als der aus einem Wortstammexperten (WSXP) und Affix tabellen in einer Transformation (genannt Wortartenexperte (WAXP)) entstehende Wortexperte die Grenze zwischen Lexikon und Regelapparat verwischt.

Trotzdem ist es möglich, den Wortstammexperten als Lexikoneintrag zu interpretieren und in diesem sprachlichen Konzept semantische Kategorien zu lokalisieren. In dem in Figur 24 gezeigten Wortstammexperten (WSXP) sind neben dem Stamm

SORTE, ARGUMENTE und SYNTAKTISCHE MERKMALE als Eintragskategorien

ausgezeichnet. Die semantischen Angaben sind in SORTE und ARGUMENTE zu finden. Die Angaben sind leichter zu deuten, wenn man die in Figur 24 benutzte Notationsweise für den WSXP-Lexikoneintrag transformiert und

"Sorte und Ergänzungsforderungen in der gleichen Struktur wie den fertigen SRL-Ausdruck notiert"
(Eimermacher83b, p.6)

Der Ausdruck nimmt dann folgende Gestalt an:

```
EREIGNIS: lieg(MO: Agens, LOC: Loc)--Syntax(V; präs; intrans; ...)
ZUSTAND: zieh(MO: Agens, DLOC: DLoc)--Syntax(V; präs; intrans; ...)
HOCH: *()--Syntax(Sort; (2.nil))
ZUSTAND: hoch(MO: *)--Syntax(Adj; pos; (*. hoeh. hoech)).
```

(Eimermacher83b, p.7)

Für das erste Beispiel ("lieg") heißen die semantischen Angaben:

- EREIGNIS: Sorte des Eintrags ("lieg ist von der Sorte EREIGNIS")
- lieg: SRL-Operator (der Stamm ist ein Operator, semantisch betrachtet)

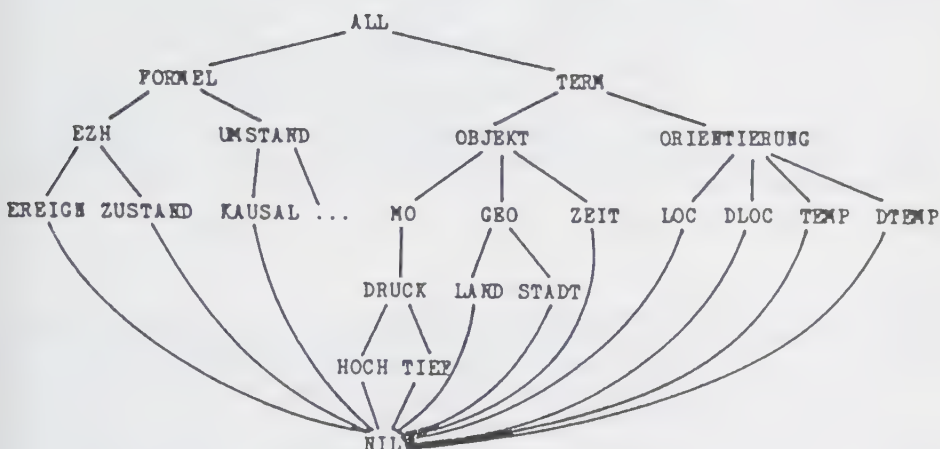
ein Prädikat mit Argumenten, vom Parsing betrachtet ein potentieller Wortexperte)

- (MO: Agens, LOC: Loc): Liste der möglichen Ergänzungen (hier sind es zwei: der Agens-Kasus ist aus dem Sortenbereich MO -meteorologisches Objekt-, der Locative-Kasus aus dem Sortenbereich LOC = Ortsangabe)

Die Ähnlichkeit der WSXP-Einträge mit SRL-Ausdrücken beruht darauf, dass sowohl WSXP-Lexikoneinträge als auch semantische Repräsentationsausdrücke SORTEN-Angaben enthalten. Diese Sortenangaben beziehen sich beim SRL-Ausdruck sowohl auf den Operator als auch auf Argumente und beim WSXP-Lexikoneintrag sowohl auf Stämme als auch auf die Ergänzungen. Die Sorten-Angaben sind in Sortenverbänden geordnet. Es sind diskursspezifische semantische Kompatibilitäten definiert. M. Eimermacher bemerkt dazu

"Wir gehen davon aus, dass die Semantik zumindest teilweise textklassenspezifisch bzw. diskursabhängig definiert werden muss (hier für Wetterberichte)"
(Eimermacher83c, p. 15)

Die folgende Figur zeigt einen vereinfachten Ausschnitt aus dem Sortenverband für Wetterberichte:



EZH= Ereign., Zustand, Handlg. GEO = Geograph: Objekt
 wobei Handlungen in LOC = Ortsangabe
 Wetterberichten fehlen DLOC= Ortsveränderung
 DRUCK = Druckgebiet TEMP= Zeitangabe
 MO = Meteorolog. Objekt DTEMP=Zeitraum, zeitl. Veränderung

Figure 27. 'Sortenverband (KIT)': (Eimermacher83c,p.14)

Zusammenfassend kann gesagt werden, dass in KIT folgende lexiko-semantische Kategorien verwendet werden:

- SORTE bezogen auf einen Eintrag (Stamm)
- TIEFENKASUS als Ergänzungen zum Stamm (falls es Ergänzungen hat)
- SORTEN zu den Ergänzungen.

Entsprechend dem Sortenverband sind auch Obersorten zugelassen.

Zu den Tiefenkasus und Ergänzungen erläutert Eimermacher in bezug auf das kommentierte Beispiel "lieg":

"MO und LOC sind sozusagen die Slots des Konzepts "lieg" auf sprachlicher Ebene."
 (Eimermacher83b,p.6)

Zur "Tiefe" der Tiefenkasus bemerkt Eimermacher:

"Als "TKs" verwenden wir nicht die "tiefste Ebene", wie sie für Übersetzungssysteme oder syntaktisch schwache Sprachen (z.B. Englisch) gefordert wird [...]. Diese Ebene hätte den Nachteil, dass sie kaum ohne Inferenzen darstellbar wäre. Wir wollen die Inferenzen zunächst von dem ersten Schritt, dem reinen Parsen, trennen und darauf verzichten,

'ich helfe DIR' und 'ich unterstütze DICH'

auf denselben TK abzubilden. (Man würde hier zwischen 'Beneficiary' und 'Patiens' trennen). Ausserdem fassen wir GOAL und SOURCE zu einem TK 'DLOC' (D=Delta: "Veränderung") zusammen und führen entsprechend TEMP und DTEMP für die Zeit ein"
(Eimermacher83b, p.13)

*

EVAR:

Dem Lexikonaufbau im System EVAR liegt die Idee der engen Kopplung von zwei semantisch relevanten Modellen zugrunde. Es sind dies

1. Das Valenzmodell (3-Stufenmodell nach Helbig/Schenkel) und
2. Die (Tiefen)-Kasustheorie (nach Fillmore).

Der Sinn dieser Kopplung wird aus folgendem Ausschnitt eines Lexikoneintrages klar:

FAKULTATIV 1

KASUSFUNKTION: Nominalgruppe (KASUS: Genitiv)

TIEFENKASUS: Object

SELECTION: Con, Res

KASUSFUNKTION: Präpositionalgruppe

(PRÄPOSITION: fuer;

KASUS: Akkusativ)

TIEFENKASUS: Object

SELECTION: Con, Res

(Ehrlich83, p.71)

U. Ehrlich kommentiert dieses Beispiel:

"Kann ein Aktant alternativ auf verschiedene Weisen aktualisiert werden, müssen für die verschiedenen Möglichkeiten der Realisierung vollständige Einträge mit Kasusfunktion, Tiefenkasus und Selektion erstellt werden". Und sie fährt fort:

"Dies heisst nicht, dass beide Möglichkeiten, also Nominalgruppe und Präpositionalgruppe, zusammen realisiert sein können, sondern jeweils nur eine von ihnen".

(Ehrlich83, p.71)

Die Aktantenrealisierungen im Sinne des Modells von Helbig dienen dazu, die entsprechende Gruppe (hier nominale oder präpositionale Gruppe) oberflächensyntaktisch zu erkennen, während im Tiefenkasus die semantische Reduktion des Syntagmas auf einen Kasus ausgedrückt ist.

Die Funktion des eben präsentierten Beispiels wird noch klarer, wenn der Aufbau des semantischen Lexikoneintrags im System EVAR nicht nur ausschnittsweise betrachtet wird, sondern ganzheitlich:

"Das Lexikon des Erlanger Spracherkennungsprojektes besteht in einem Vollformenlexikon. Es werden also alle Flexionsformen eines Wortes als gesonderte Worteinträge ins Lexikon aufgenommen. [...] Da die semantischen Eigenschaften bei jeder Form eines Wortes gleich sind, muss der semantische Lexikon-Eintrag nur einmal bei der Grundform eines Wortes erstellt werden.

In den semantischen Teil des Lexikon-Eintrages gehen alle Bedeutungsvarianten eines Wortes ein. [...] Alle möglichen Bedeutungsvarianten werden so nummeriert hintereinander aufgeführt.

Für jede Bedeutung wird das Wort einer semantischen Klasse zugeordnet."
(Ehrlich83, p.70)

An die derart auf die entsprechenden Vollformen bezogene und in Bedeutungsvarianten differenzierte GRUNDFORM wird für jede VARIANTE der entsprechende Valenzrahmen, "bzw. die semantische Funktion, die das Wort im Satz oder einer Wortgruppe übernimmt" (Ehrlich83, p.70), aufgesetzt. Ein Valenzrahmen wird "bei Verben, Substantiven und attributiv oder prädikativ gebrauchten Adjektiven" verwendet, die semantische Funktion bei "Adverbien und adverbial verwendeten Adjektiven".
(alle drei Zitate: Ehrlich83, p.70)

Aus dieser ganzheitlichen Perspektive können die wesentlichen Leistungen des Lexikons wie folgt bestimmt werden:

1. Formen-Lemma-Verknüpfung,
2. semantische Variantendifferenzierung,
3. Valenz- und Funktionszuordnungen

Besondere Beachtung verdient der dritte Punkt, weil in dieser Formulierung der Valenzbegriff so weit gefasst ist, dass er weder auf eine einzige Wortart beschränkt, noch einseitig syntaktisch oder semantisch bewertet wird.

Die bei diesem Ansatz benötigten semantischen Merkmale bilden hierarchische Merkmalordnungen, die teilweise domain-unabhängig sind, teilweise dem Diskursbereich "Intercity" angepasst wurde. Die Merkmalordnungen dienen als Wertebereiche für die Klassifizierung der Lemmata (KLASSE) und der Aktanten (SELEKTION), wobei an einer Stelle mehrere Merkmale gesetzt werden können.

Die im Lexikon verwendeten semantischen Kategorien sind folgende:

- BEDEUTUNGSVARIANTE, wobei man sagen kann, dass die Variante sich bezüglich einer Grundform wie eine Kategorie verhält, aber selber auch Kategorienträger ist. (KLASSE, VALENZ).
- KLASSE (klassiert wird die Grundform, die Bedeutungsvariante, die Klassenordnungen sind wortartenspezifisch).
- TIEFENKASUS (auch als "logische Valenz" verstanden).

- SELEKTION (des Kasus).

Die Kategorie Kasusfunktion ist eine morphosyntaktische Kategorie.

*

SADAW:

Im Wörterbuch SADAW sind vorwiegend morphosyntaktische Angaben codiert. Allenfalls liessen sich hier semantische Kategorien aus dem ebenfalls im SFB100 verwendeten Lexikon DEUSEM nennen:

- Relationen (Synonymie, Hypernymie, Antonymie, Inferenz)
- semantische Merkmale (Abstrakt, Belebt, Instrument u.a.m.)
- semantische Merkmale speziell für Präpositionen, basierend auf einer Nomenklassenliste (Nomina mit möglicher lokaler Bedeutung, Nomina zu Bezeichnung von Zeitintervallen u.a.m.)

Diese Relationen und Merkmale sind bei (Hess83, p.177) zusammengestellt.

Der Einbezug dieser semantischen Kategorien würde jedoch den durch die Antworten im Fragebogen gegebenen Rahmen sprengen, da dort nicht auf das Lexikon DEUSEM verwiesen wurde. (In Anmerkung [**9] zu Teil 3] sind jedoch einige Angaben zu DEUSEM zu finden).

Aus demselben Grund werden auch neuere semantische Modelle aus dem SFB100 nicht einbezogen [**8].

*

ALEXSYS:

Im kumulierten Lexikon sind die semantischen Kategorien in einer eigenen Informationsklasse zusammengefasst. In (Hess83) werden auf Seite 201 folgende genannt, wobei jeweils vermerkt ist, in welchem (welchen) Einzelllexikon (Einzelllexika) die meisten Belegungen der entsprechenden Kategorie zu finden sind:

- Semantische Definition (Wahrig)
- Semantisch relevanter Kontext (Wahrig)
- Anwendungsbeispiele (Wahrig)
- Semantische Relationen (HAM/RPM; BACON; DEUSEM; JUDO) [**9]
- Semantische Merkmale (BACON; DEUSEM; HAM/RPM; JUDO; LIMAS; PLIDIS)
- Fachbereichskennzeichen (SYSTRAN)

Ebenfalls bei (Hess83) findet sich auf Seite 219 noch die zusätzliche Kategorie

- Semantische Primitiva.

Teilweise könnten noch die in der Informationsklasse "Syntaktische Angaben" unter "Tiefenkasus" verzeichneten Kategorien in Frage kommen:

- Subjekt-/Objekt-Code für Verben (SYSTRAN)
- grammatische Inhalte von Präpositionalphrasen (PRÄPLEX)
- mögliche Nominalkomplexe in Verbindung mit Präpositionen (PRÄPLEX)

Die Problematik der semantischen Kategorien bei ALEXSYS besteht darin, dass diese Kategorien nur schwach belegt sind und dass die

vorausgesetzten Semantikkonzeptionen in denjenigen Einzelexika, welche semantische Kategorien aufweisen, stark divergieren. (Vergl. dazu Anmerkung [**9] zu Teil 3)).

*

TEAM:

Von den 10 Kategorien im Eintragskopf (vergl. Fig. 18 u. 19) können die folgenden als semantisch bezeichnet werden:

- 03 Bewertung der Äquivalenzbeziehung ("Arbeitsbenennung" oder definierter, gesicherter Äquivalenzgrad)
(Einschränkend muss hier bemerkt werden, dass es sich bei dieser Kategorie und den beiden folgenden teilweise um pragmatische Angaben handelt)
- 06 Fachgebietskennung
- 09 Geltungsbereichskennung

Von den 10 Kategorien im Sprachblock "Deutsch" können folgende als semantisch bezeichnet werden:

- 10 Benennung oder fachsprachliche Wendung, ev. mit klärendem Zusatz.
- 14 Definition, Erläuterung (jeweils mit Quellenangabe)
- 15 Anwendungsbeispiel
- 16 Synonym(e)
- 17 Suchwort (-wörter) zur Benennung bzw. Wendung in Feld 10.

Die Suche nach semantischen Kategorien in einer mehrsprachigen Terminologiedatenbank ist nur sinnvoll, wenn man einen fachsprachbezogenen Semantikbegriff verwendet. Deshalb ist speziell auf die oben genannte Kategorie

"Benennung oder fachsprachliche Wendung"

hinzuweisen. "Benennung" ist hier exakt definiert und im Sinne der DIN 2330 verwendet (Vergl. Anmerkung [**1] in Teil 1). In dieser Norm finden sich Anleitungen und Grundsätze zur Begriffsbildung, Merkmalszuweisung, Definition und Vermeidung von Mehrdeutigkeiten.

*

3.2.2 Einbeziehung des semantischen Bezugssystems.

Um die Kategorien vergleichend zu analysieren, empfiehlt es sich, die den Lexika zugeordneten semantischen Bezugssysteme im Sprachverarbeitungskontext näher zu betrachten. Der Begriff "semantisches Bezugssystem" soll dabei zweierlei ausdrücken:

1. Im engeren Sinn ist darunter die im System verwendete semantische Repräsentation zu verstehen (falls vorhanden) und
2. im weiteren Sinn des Begriffs der semantisch-pragmatische Kontext ("Background") des Systems.

Im Sinne einer von Primio und Christaller vorgenommenen terminologischen Klärung kann anstatt von 'FLAVOR' (für generisches Objekt im Kontext Objekt-orientierter Programmierung) auch von 'CLASS' gesprochen werden (vergl. Primio83, p.6).

Einen Eindruck von der etwas eigenwilligen Terminologie vermittelt folgende Stelle bei (Primio83):

"Flavors are organized in a directed graph called the 'flavor graph'. In most applications this will be a tree, in which case the root is the 'vanilla flavor'. It corresponds to the 'thing frame' in FRL. The flavor graph provides the 'heritage' of information 'downwards'. When defining a new flavor it must be specified where in the graph it should fit in. This is done by naming the parent nodes, which are called the 'superclasses' of the new flavor. [...] The new flavor will inherit all information contained in its superclasses, and normally will add new information given in its definition which distinguish it from all other flavors"

(Primio83, p.6)

"The information that a flavor holds can be divided into two classes: information owned by the flavor itself and inherited information, i.e. that is owned by all superclasses of the flavor, its super-superclasses, etc. Owned information might modify or even override the inherited information."

(Primio83, p.6)

Abgesehen von der Unterscheidung "geerbter" und "eigener" Information innerhalb eines "flavor", wird zusätzlich unterschieden, ob die enthaltene Information prozeduralen oder deklarativen Charakter hat:

"What might be called the 'declarative knowledge' of a flavor is stored in 'variables' of different kinds. The 'procedural' knowledge is encapsulated in so called 'methods'."

(Primio83, p.7)

Da flavor's Objekte im Sinne der Terminologie von objekt-orientierter Programmierung (z.B. SMALLTALK) sind, können zwischen ihnen Botschaften (messages) getauscht werden. Die Objekte reagieren darauf einerseits wie Programme auf Parameter, aber andererseits werden sie wie Datenstrukturen benutzt, denn

"objects behave sometimes like programs or procedures and at other times like ordinary data structures".

(Primio83, p.6)

*

Die semantischen Bezugssysteme der folgenden 5 Systeme sind weniger heterogen und werden lediglich benannt, da der Kontext aus vorangehenden Passagen vorliegender Arbeit hergestellt werden kann.

KIT: Logische Prädikate und die logisch/semantische Repräsentationssprache SRL (Habel/Reddig/Rollinger81).

EVAR: KL-ONE-ähnliche Netze (Brietzmann84)
(Brachmann77).

SADAW: Maschinelle Übersetzung sowie Information und Dokumentation (CTX) (Zimmermann84) (Weber84).

ALEXSYS: Generelle Wortdatenbank ohne Fixierung auf ein bestimmtes semantisches Bezugssystem.

TEAM: Fachsprachliche Definitionen, Beispiele, Übersetzungshilfe, terminologische Normung.

*

3.2.3 Vergleich ableitbarer semantischen Konzeptionen.

Um einen Überblick zu verschaffen, werden nun in einer Tabelle die bereits erwähnten semantischen Kategorien und ihre Bezugssysteme summarisch zusammengestellt.

HAM	KIT	EVAR	SADAW	ALEXSYS	TEAM
LEMMA SYNONYM PARA- PHRASE KLASSE KASUS- RAHMEN ROLLEN- RESTRIK- TION	STAMM LESART SORTE TIEFEN KASUS SORTE DER TK(s)	GRUNDFORM BE- DEUTUNGS- VARIANTE KLASSE VALENZ- RAHMEN SELEKTION DER TK(s)	IDENTIFI- KATIONS- WORTLAUT PLUS BE- DEUTUNGS- NUMMER	LEMMA DEFINITIO KONTEXT BEISPIEL RELATION MERKMALE FACHGEB. PRIMITIVA	TERM (ÄQUIVA- LENZ) (FACHGEB) (GELTUNG) BENENNUNG DEFINITIO BEISPIEL SYNONYM SUCHWORT
SURF	SRL	KL-ONE	ANALYSE- 'LEVEL'	KUMULA- TION	TERM- GLEICHUNG

An dieser Tabelle lassen sich folgende für den Vergleich relevante Beobachtungen machen:

1. In der Systematik der Kategorien und auch in der Benennung der einzelnen lexiko-semantischen Kategorien sind Ähnlichkeiten zwischen den Kategorien in den Kolonnen unter HAM, KIT und EVAR festzustellen. Dies betrifft die folgenden Kategorienreihen:

- KLASSE	SORTE	KLASSE
- KASUS- RAHMEN	TIEFEN- KASUS	VALENZ- RAHMEN
- ROLLEN- RESTRIK- TION	SORTE DER TK(s)	SELEKTION DER TK(s)

Obwohl es falsch wäre, die Differenzen zu ignorieren, kann bei diesen drei Kategorien-Reihen von starker Ähnlichkeit gesprochen werden, die auf die Verwendung ähnlicher Modellvorstellungen im Bereich der Semantik hindeutet.

Indirekt wird diese Vermutung dadurch bestätigt, dass in der untersten Zeile der Tabelle, wo die semantischen Bezugssysteme aufgeführt sind, bei diesen drei Lexikonsystemen, im Unterschied zu den übrigen, künstliche semantische Repräsentationssprachen notiert sind.

2. Kontrastierend zu diesen Reihen können Ähnlichkeitspaare unter den Kolonnen ALEXSYS und TEAM gebildet werden:

- DEFINITION DEFINITION
- BEISPIEL BEISPIEL
- FACHGEBIET (FACHGEBIET)

Obwohl diese Kategorien in ihrer Benennung identisch sind, bedeutet dies natürlich nicht, dass sie auch sonst vollkommen identisch sind.

3. Eine andere Gemeinsamkeit verbindet sämtliche Kolonnen über die Kategorienreihe in der ersten Zeile der Tabelle: In allen sechs Lexikonsystemen werden die semantischen Kategorien auf eine von der BEDEUTUNG her ausgezeichnete Wortform bezogen. Interessant ist, dass dies auch für SADAW zutrifft, obwohl sich dieses Wörterbuch von seinen übrigen Kategorien her als morphosyntaktisch bezeichnen lässt. Aber auch bei dieser Beobachtung wäre es unvorsichtig, die Differenzen zwischen den einzelnen Lemma-Begriffen zu vernachlässigen. Trotz solcher Differenzen sind die eben vorgenommenen Ähnlichkeits-Gruppierungen in der Tabelle der semantischen Kategorien nicht zufällig zustande gekommen:

Man könnte die Ähnlichkeiten (1) und (2) derart interpretieren, dass die Reihen unter HAM, KIT, EVAR und diejenigen unter ALEXSYS und TEAM auf zwei verschiedene Semantikansätze hindeuten. Als Hypothese kann formuliert werden, dass in den drei Systemen aus dem Umfeld der Künstlichen Intelligenz Semantikmodelle mit der Fähigkeit zur Transformation von morpho-syntaktischen Strukturen in semantische und pragmatische Repräsentationen verwendet werden, und dass in den beiden Systemen ALEXSYS und TEAM solche Modelle fehlen. Auch für ALEXSYS und TEAM kann die Hypothese positiv formuliert werden: Die zugrundeliegende Semantikvorstellung in den beiden Systemen ALEXSYS und TEAM ist weniger an Transformationen als vielmehr an Äquivalenzrelationen in Gestalt von Definitionen und Paraphrasen orientiert.

Das Wörterbuch SADAW kann weder der einen noch der anderen Gruppe zugeordnet werden. Bei isolierter Betrachtung dieses Lexikons ist schwer auszumachen, welche Semantikvorstellung zugrundeliegt. Deutlicher wird dies jedoch, wenn SADAW im Applikationskontext "M.Ü." oder "I.u.D" oder "automatische Lemmatisierung" betrachtet wird. Am Beispiel von Information und Dokumentation (I.u.D) soll das gezeigt werden:

Für das System JUDO (Juristische Dokumentenanalyse, H. Zimmermann) stellt SADAW lediglich die erste Stufe in einer ganzen "Kaskade"

von Wörterbüchern (und Analyseprozessen) dar:

- Syntaxwörterbuch SADAW
- Semantikwörterbuch DEUSEM
- Fachspezifisches Wörterbuch FAWO

sowie FALEX, SEMREL und INFO.
(Werner80, p.90)

Der Zusammenhang zwischen den drei Komponenten FAWO, FALEX und SEMREL ist in dieser Applikation folgender:

"DEVIL [Deskriptorerstellungsverfahren, H.H.] greift auf das fachgebietsspezifische Wörterbuch FAWO und die Thesaurusrelationen SEMREL nicht direkt zu, sondern über das Format FALEX (fachspezifisches Lexikon), in dem die Information von FAWO und SEMREL gemeinsam vorhanden sind."
(Werner80, p.98)

DEUSEM enthält "semantische Merkmale von Substantiven und Regeln zur Erkennung fester Wendungen und zur Vereindeutigung mehrdeutiger Wörter" (Werner80, p.90). (vergl. Anmerkung [**9] zu Teil 3).

FAWO enthält "Angaben über die Fachgebietszugehörigkeit und Relevanz der Lemmanamen" (Werner80, p.90).

SEMREL enthält Begriffspaare mit ihren Begriffsbeziehungen (Werner80, p.91).

INFO wird von Werner wie folgt beschrieben:

"INFO = (Informationsdokumente über die Bedeutungs differenzierung von Homonymen) mit definierenden und beschreibenden Erklärungen der Bedeutungsvarianten von Homonymen, ergänzt durch Gewichtungangaben, die eine Schätzung der Häufigkeit der Bedeutungsvarianten im Fachgebiet darstellen." (p.91)

Man könnte die hier zugrundeliegende Semantikvorstellung als "funktional" bezeichnen, indem semantische Information applikationsspezifisch variabel gehalten ist. Im Falle von JUDO wären das vor allem die fachgebietsspezifischen Relevanzangaben in FAWO.

Für einen anderen Applikationszusammenhang, z.B. "Maschinelle Übersetzung" (M.Ü., SUSY, H.-D. Maas) werden die in SADAW gehaltenen Informationen zwar beibehalten (inklusive der nicht nur syntaktisch-distributional, sondern auch semantisch motivierten Eintragsidentifikation über Identifikationswortlaut und Bedeutungsnummer), aber die applikationsspezifischen Semantikwörterbücher können ausgewechselt, bzw. umformatiert werden. (Der in der oben angeführten Überblickstabelle unter SADAW als semantisches Bezugssystem angegebene Begriff "Analyse-level" soll auf die Variabilität und Funktionalität der zugrundeliegenden Semantikkonzeption hindeuten, welche aber auf einer Analyseebene zu lokalisieren ist, die nicht mithilfe von SADAW bearbeitet wird).

*

Die vorgenommene grobe Unterscheidung der drei Semantikvorstellungen

- "Transformation in eine Repräsentation" (a)

- "Definition und Paraphrase" (b)

- "Applikationsvariable Semantik" (c)

ist nicht zu verwechseln mit den in Kapitel (2.1.*) erwähnten SemantikMODELLEN, im Sinne der Terminologie von F. Guenther und H. Lehmann, denn die dort vorgenommene Unterscheidung zwischen dem "logischen" und dem "repräsentationellen" Modell ist feiner als die eben getroffene, denn jene Modelle lassen sich beide der Vorstellung (a) zuordnen.

Auf der jetzt diskutierten abstrakteren Ebene des Vergleichs der Semantikkonzeptionen soll speziell auf die Differenz zwischen den Vorstellungen (a) und (b) eingegangen werden. Vorangehend aber noch eine Bemerkung:

Die hier gebildeten Gruppen sind nicht als Einheiten oder Modelle zu betrachten, denn die Modelle sind in den konzeptionell durchaus unterschiedlichen Lexikonsystemen selbst zu finden und nicht in den abstrakteren zugrundeliegenden Vorstellungen, von denen jetzt die Rede ist. Deshalb sollte es auch nicht erstaunen, die beiden auf den ersten Blick gänzlich verschiedenen Lexikonsysteme ALEXSYS und TEAM in einer Gruppe zu finden.

Die Zuordnung des Systems ALEXSYS zur Gruppe (b) basiert auf folgenden Überlegungen:

- Die drei semantischen Kategorien "semantische Definition", "semantisch relevanter Kontext" und "Anwendungsbeispiele" lassen sich problemlos der Gruppe "Definition und Paraphrase" zuordnen, sind doch diese Kategorien an den im Einzelllexikon "Wahrig" verwendeten definitorischen Paraphrasen orientiert.

- Die Kategorien "semantische Relationen" und "semantische Merkmale" sind wesentlich heterogener belegt. Sie entstammen u.a. den nicht sehr umfangreichen Lexika aus folgenden Projektzusammenhängen: HAM/RPM, BACON, PPLIDIS (Vergl. Anmerkung [**9] zu Teil 3). Diese Projekte waren alle K.I.-orientiert, mit jeweils sehr unterschiedlichen Anwendungsgebieten:

PLIDIS: Abwasserkontrolle,
HAM/RPM: Hotelzimmerreservation,
BACON: Frage-Antwort-System.

In die gleichen Kategorien "semantische Merkmale" und "semantische Relationen" fanden auch Relationen und Merkmale aus den Lexika DEUSEM und aus den in JUDO gebrauchten Wörterbüchern (INFO) Eingang.

- Aufgrund der Heterogenität der Einträge in den Kategorien "semantische Relationen" und "semantische Merkmale" ist es angebracht, die hier zugrunde liegende Semantikvorstellung ebenfalls mit "Paraphrase" zu umschreiben, allerdings nicht im Sinne der traditionellen lexikographischen Paraphrase (verstanden als Explikation), sondern als Umschreibung des Prozesses "Kumulation" von Semantik-Information aus heterogenen Einzelllexika in übergeordnete kompatible Kategorien.

- Gemäss der Konzeption "Allgemeine Wortdatenbank" ist ALEXSYS in bezug auf die Semantikategorien weder als Transformationskomponente (Semantikvorstellung (a)), noch als Basis für Funktionen in applikationsvariablen Zusammenhängen (Vorstellung (c)) konzipiert, sondern als Konsultationslexikon für Linguisten.

Gerade in diesem letztgenannten konzeptionellen Zusammenhang überschneidet sich die Semantikvorstellung bei ALEXSYS mit derjenigen bei TEAM:

Auch bei TEAM geht es hauptsächlich um die Konsultation einer Datensammlung für spezifische Benutzerkreise (Übersetzer, Terminologen und Normspezialisten). Diese Funktion des Lexikongebrauchs entspricht der traditionellen Vorstellung vom Nutzen eines Lexikons. Dass bei TEAM auch noch andere Funktionen für den Lexikongebrauch vorgesehen sind, wie die Wörterbuch- und Glossargenerierung, ändert nichts an der Tatsache, dass das auf terminologischen Gleichungen aufbauende und mit fachlichen Definitionen und Benennungen versehene System eindeutig der Gruppe (b) "Definition und Paraphrase" zugeordnet werden kann.

Die Diskussion der Differenz zwischen den Semantikvorstellungen der Gruppen "Transformation in eine Repräsentation" (a) und "Definition und Paraphrase" (b) nimmt damit eine neue Dimension an, indem gesagt werden kann, dass die Gruppe (b) eher dem traditionellen Typ eines Lexikons entspricht und die Gruppe (a) eine Lexikonkonzeption repräsentiert, wie sie bis vor wenigen Jahren überhaupt nicht existierte.

In der traditionellen Lexikographie wird vorwiegend Typ (b) reflektiert (Vergl. u.a. (Hausmann85), (Lutzeier85), (Drosdowski77), (Henne79), (Wolski82), (Zgusta85)).

Es gibt allerdings lexikographische Ansätze, die, aufbauend auf den Konzepten zu "lexikographischer Definition" (Vergl. zu "Definition" auch Anmerkung [**17] zu Teil 1), eine "Modernisierung" der Paraphrase und der Definition in Definitionswörterbüchern (einsprachigen Bedeutungswörterbüchern) anstreben. Als Beispiel dafür kann die von F. Neubauer 1980 durchgeführte vergleichende lexiko-semantische Analyse deutscher einsprachiger Wörterbücher genannt werden (Neubauer80b).

Für ein besseres Verständnis der Differenz zwischen lexikalischen Semantikkonzeptionen des Typs (a) und des Typs (b) lohnt es sich im Rahmen unseres Vergleichs semantischer Kategorien und semantischer Bezugssysteme maschinenlesbarer Wörterbücher, die von Neubauer geleistete Vergleichsarbeit bezüglich gedruckter einsprachiger deutscher Wörterbücher näher zu betrachten. Es geht dabei hauptsächlich um den Zusammenhang der Begriffe "Paraphrase", "Definition" und "Raster".

Die definitorische Paraphrase eines Begriffs kann derart erweitert werden, dass sich die Definition als eine Abbildung des zu definierenden Begriffs in ein systematisch aufgebautes Paraphrasemuster beschreiben lässt. In (Neubauer80b) wird exemplarisch gezeigt, wie derartige Paraphrasemuster (Raster) durch Vergleich von Informationskategorien in Wörterbuchartikeln aus verschiedenen Wörterbüchern zum selben Lemma extrahiert werden können.

Neubauer differenziert dabei zwischen den verschiedenen Kategorisierungsaspekten des gleichen Lemmas. Werden z.B. Lemmata aus dem Bereich ARBEIT paraphrasiert, kann dies sowohl unter arbeitsrechtlich motivierten Gesichtspunkten als auch unter dem Aspekt von Berufsbezeichnungen geschehen. Arbeitsrechtlich motivierte Explikationen sind etwa bei den Begriffen ARBEITER, ANGESTELLTER und BEAMTER angebracht.

Für die Lemmata 'Arbeiter', 'Angestellter', 'Beamter' in den Wörterbüchern

- Klappenbach/Steinitz (hg), Wörterbuch der deutschen Gegenwartssprache, 6 Bde, Berlin, 1961-77
- Drosdowski et al., Duden: Das grosse Wörterbuch der deutschen Sprache in sechs Bänden, Mannheim, 1976ff
- Drosdowski et al., Duden: Stilwörterbuch der deutschen Sprache (= Der grosse Duden, Bd. 2), Mannheim, 1971
- Wahrig, Deutsches Wörterbuch, Gütersloh, 1975
- Köster et al., Ullstein-Lexikon der deutschen Sprache, Frankfurt, 1969.

extrahiert Neubauer den vorgefundenen Informationskategorien den folgenden Raster für den Standardartikel für ARBEITER und ANGESTELLTER:

"jemand (Proform), der eine (überwiegend) körperlich/geistige Tätigkeit in ... (Arbeitsort) ausübt, durch ... bezahlt wird, im Alter Anspruch auf eine Rente/Pension (Altersversorgung) hat und dessen Arbeitsverhältnis beidseitig gekündigt werden kann / auf Lebenszeit läuft."

(Neubauer80, p.109)

"Angewandt auf ARBEITER lautet dieser Explikationsversuch dann:

jemand, der eine (überwiegend) körperliche Tätigkeit auf dem Lande, in der Industrie oder im Gewerbe ausübt, durch Stunden- oder Leistungslohn bezahlt wird, im Alter Anspruch auf eine Rente hat und dessen Arbeitsverhältnis beidseitig gekündigt werden kann.

Bei BEAMTER erübrigt sich Information über Art der Tätigkeit, Bezahlung, da diese über den Verweis auf ANGESTELLTER erreichbar ist:

jemand, der als Angestellter im öffentlichen Dienst tätig ist, im Alter Anspruch auf eine Pension hat und dessen Arbeitsverhältnis auf Lebenszeit läuft."

(Neubauer80, p.109)

Für die Begriffe ARZT, BAUER, BUCHDRUCKER, DREHER extrahiert Neubauer den

"Raster für den Standardartikel für Berufsbezeichnungen":

"jmd. (proform), der als (Fach)Arbeiter / Angestellter / Selbständiger arbeitet, seine ... (Ausbildung) abgeschlossen hat / ohne Ausbildung ist und in ... (Ort) mit ... (Instrumente/Hilfsmitteln) die ... (Tätigkeit) von/an ... (Personen/Objekten) durchführt (wodurch ... (Ergebnis der Tätigkeit))."

(Neubauer80, p.110/111)

"Angewandt auf Arzt lautet dieser Explikationsversuch dann:

jmd. der als Selbständiger arbeitet, sein Medizinstudium abgeschlossen und eine staatliche Zulassung erhalten hat und in einer Praxis oder Krankenhaus mit Medikamenten und Geräten die Behandlung von Kranken durchführt, (wodurch diese gesund werden" [sollen H.H.]).

(Neubauer80, p.111)

Neubauer kommt in Bezug auf die 6 von ihm analysierten Wörterbücher zu kritischen Feststellungen in bezug auf die Einheitlichkeit und die Konsistenz der in den Wörterbuchartikeln durchgeführten Paraphrasierungen, sowohl hinsichtlich der Verweistruktur zwischen den onomasiologisch zusammenhängenden Begriffen als auch hinsichtlich von Lücken innerhalb der Wörterbuchartikel. Einige Beispiele mögen das illustrieren:

"Auffallend ist auch das Informationsgefälle zwischen ARZT und BEAMTER im Vergleich mit BUCHDRUCKER und DREHER. So wird die Ausbildung des Arztes und seine staatliche Zulassung in allen Wörterbüchern ausführlich geschildert, Lehre und Gesellenprüfung in handwerklichen Berufen aber unerwähnt gelassen. Mit Ausnahme der Drehbank werden keine Instrumente, Maschinen oder sonstige Mittel angegeben, mit denen die Tätigkeit ausgeübt werden könnte. Über die eigentliche Tätigkeit des Beamten erfährt man nichts, sein wichtigstes Merkmal scheint die Pensionsberechtigung zu sein, diese trifft aber auch auf Abgeordnete zu. D6 [= der sechsbändige Duden, H.H.] kennt Bundes-, Landes und Gemeindebeamte, U [= Ullstein-Lexikon] nur Bundes- und Landesbeamte. W [= Wahrig] auch Privatbeamte. Die Altersversorgung von Arbeitern und Angestellten ist nur in K [= Klappenbach] erwähnt. GEHALT führt in den meisten Wörterbüchern zu Zirkularität: In W ist ein Angestellter 'jemand, der gegen ein festes Gehalt ... arbeitet', ein Gehalt dagegen ist 'das Arbeitsentgelt für Beamte und Angestellte'."

(Neubauer80, p.112)

Neubauers Kritik beruht auf vielen konkreten Beispielen, die alle nach dem gleichen Verfahren abgehandelt werden:

1. Vergleich der 6 Wörterbücher.
2. Informationsauswahl.
3. Ableitung eines allgemeinen Rasters.
4. Anwendung des Rasters auf konkrete Fälle.
5. Kritische Gegenüberstellung der in den Wörterbuchartikeln durchgeführten Paraphrasierungen.

Um in einem Bedeutungswörterbuch eher zu einheitlichen und konsistenten Paraphrasierungen zu kommen, könn(t)en von Lexikographen auch begriffliche Strukturbilder als onomasiologische

Orientierungshilfen benutzt werden.

Gerade auch zum Gebiet ARBEIT bringt E. Wiegand in 'Nachdenken über Wörterbücher: Aktuelle Probleme' (in: Drosdowski79) einen Strukturbildausschnitt als onomasiologische Orientierungshilfe bei der lexikalischen Paraphrasierung.

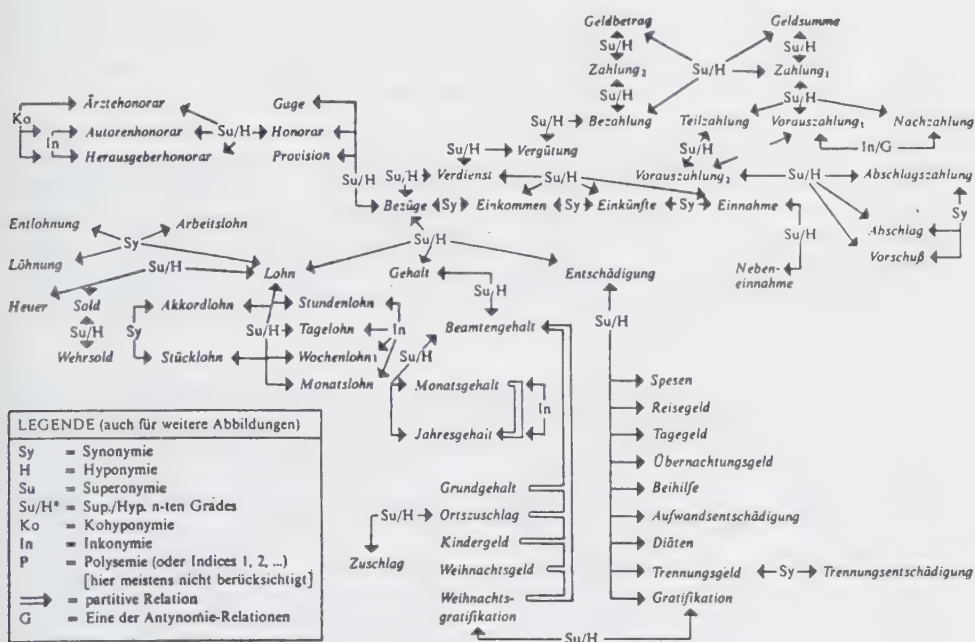


Figure 28. 'Onomasiologisches Strukturbild': (Wiegand in Drosdowski77, p. 86)

*

Es wäre einseitig, beim Feststellen der Differenz zwischen den Semantikkonzeptionen

- "Transformation in eine Repräsentation" (a) und

- "Definition und Paraphrase" (b)

nur die Konzeption (b) ausführlich darzustellen, wie das eben durch näheres Eingehen auf (Neubauer80b) geschehen ist, und die Konzeption (a) nur summarisch zu behandeln.

Allerdings kann die Erläuterung der Konzeption (a) wesentlich knapper ausfallen, da sie zu einem guten Teil in ihren grundsätzlichen Aspekten schon in Kapitel (2.1.*) als Diskussion der Positionen "Repräsentationeller Ansatz" und "Logischer Ansatz" vorgestellt wurde. Auch die in Figur 4 (Teil 1 der Arbeit) abgebildeten "Wissensquellen und Analyseprozesse" nach Winograd zeigen, worum es geht:

In (logisch) getrennten Analyseschritten sollen mithilfe von den diesen Schritten zugeordneten Wissensquellen Sprachäusserungen in künstliche Repräsentationen transformiert werden, um sodann auf diesen Repräsentationen Inferenzen machen zu können.

Für die drei untersuchten Systeme HAM, KIT und EVAR ist in diesem Transformationsprozess die Kasusgrammatik besonders relevant (vergl. dazu die semantischen Kategorien). In diesen Systemen handelt es sich dabei nicht unbedingt um die reine Verwirklichung der "klassischen" Kasusgrammatik von Charles Fillmore, sondern eher um eine auf die praktischen Bedürfnisse im jeweiligen System zugeschnittene (operationalisierte) Version. Ohne auf die Besonderheiten der Verfahren in ihren einzelnen Ausprägungen einzugehen, kann der dabei stattfindende Transformationsprozess schematisch so charakterisiert werden:

1. Sätze können so geparkt werden, dass die resultierende Struktur Konstituenten und Abhängigkeiten abbildet.
2. Die erzeugten Strukturen können mit (Tiefen)-Kasusrahmen so 'gematched' werden, dass instantiierte Kasusrahmen entstehen.
3. Die instantiierten Kasusrahmen und die noch freien Strukturen sind genügend klar, um in ein logisches oder pragmatisches Modell integriert werden zu können.

Auf dieser Abstraktionsebene betrachtet, kommt das eben geschilderte Verfahren in allen drei Systemen (HAM, KIT, EVAR) zur Anwendung. Es kann deshalb als ein wesentlicher Teil der Semantikvorstellungen dieser Gruppe gelten. Damit ist nicht gesagt, dass sich diese Vorstellungen in dem einen Verfahren erschöpfen, aber zur Differenzierung der Gruppen (a) und (b) ist es hinreichend.

Weitergehende Differenzierungen innerhalb der Gruppe (a) könnten nach verschiedenen Kriterien durchgeführt werden. Einige davon seien genannt:

- Verwendete Typen von Kasusrahmen.
- Wortarten, die als Träger von Rahmen figurieren.
- Zahl und Art der Kasus.
- Sorten, Merkmale, Klassen, die den Rahmenträgern und den Kasusrollen zugewiesen werden können.

Die eben aufgezählten Kriterien beziehen sich unmittelbar auf das Kasusgrammatische Modell. Zum Vergleich könnten aber auch Kriterien herangezogen werden, die sich auf die der Kasusanalyse vorgeordnete Syntaxvorstellung beziehen oder auf die nachgeordneten Wissensrepräsentationen, denn bei beiden ist Semantik mindestens teilweise mitgedacht. Ein derart detaillierter Vergleich würde aber zu weit führen und, abgesehen davon, im Rahmen des hier geleisteten Vergleichs der zugrundeliegenden (abstrakteren) Semantikkonzeptionen wenig bringen. Anstelle eines derartigen Detail-Vergleichs semantischer Modelle sollen in den nächsten Kapiteln noch zwei andere wichtige Vergleichsaspekte zur Sprache kommen.

3.3 * WEITERE AUSWERTUNGEN.

Eine wichtige Dimension der Auswertung, die zeitliche, kam bisher noch nicht zur Sprache. Diese Dimension lässt sich in den synchronischen und den diachronischen Aspekt einteilen.

3.3.1 Synchronische Aspekte der Umfrage.

Die Umfrageergebnisse zeigen im synoptischen Vergleich die Ausbaustufe der Lexikonsysteme für den Zeitabschnitt zwischen April 1984 und März 1985. Die Daten zu den drei KI-Systemen wurden alle gleichzeitig erhoben (März 1985), diejenigen zu den grossen Lexika SADAU, ALEXSYS und TEAM früher (zwischen April 1984 und Juli 1984). Trotz dieser Zeitdifferenz ist es möglich, von einem synchronischen Schnitt zu sprechen.

Für die drei kleineren Lexika trifft das absolut zu, für die drei grösseren nahezu, denn derart umfangreiche Datenbestände, wie sie in SADAU, ALEXSYS und TEAM verwaltet werden, besitzen eine gewisse Resistenz gegen schnelle Veränderungen.

Allerdings ist auffällig, wie häufig in den Fragebogen zu den erteilten Antworten die Einschränkungen "zur Zeit", "derzeit", "in Arbeit", "zukünftig", "momentan" u.a. verwendet wurden.

Bei den 3 grossen Lexikonsystem betrifft dies folgende Aussagen:

- SADAU:

Im Fragebogen wird zwar keine derartige Formulierung gebraucht, aber im beigelegten Bericht (Klaus-Dirk Schmitz, Struktur und Updating des Analysewörterbuchs) weist H.-D. Maas innerhalb von 4 Druckseiten gleich 4 mal mit Randbemerkungen "alt!" auf Neuerungen hin, welche

a) die Indexorganisation des Wörterbuchs betreffen (einstufig --> zweistufig), sowie

b) den Umfang des Lexikons (Zahl der Einträge).

- ALEXSYS:

a) Frage Nr. 6 (File oder Datenbank?)

Die Antwort heisst zwar fileorientiert, aber mit Zusatz: "wird als Datenbank am Institut für Deutsche Sprache in Mannheim aufgebaut"

b) Frage Nr. 17 (Benutzung des Lexikons durch welche Systeme?)

"Da als Datenbank konzipiert, sind z. Zt. keine LDV-Systeme direkt an das Lexikon angeschlossen"

c) Frage Nr. 18 (Schnittstelle zur DIN 2341?)

"z.Zt. nicht".

- TEAM:

a) Frage Nr. 13 (Wie gross ist Ihr Wörterbuch?)

Mit Schreibmaschine geschrieben steht die Antwort ca. 1,5 Mio. Einträge, aber handschriftlich fügte E. Hanis hinzu: "bis 2.0 Mio"

b) Frage Nr. 18 (Schnittstelle zur DIN 2341?)

"ist derzeit in Arbeit"

*

Diese zeitlichen Einschränkungen betreffen nicht nur Zusätze oder Details, sondern zentrale Aspekte wie

- "File/Datenbank",
- "Indexorganisation" (Zugriffsorganisation),
- Umfangzuwachs im grossen Stil und
- Kopplung mit anderen LDV-Systemen.

Daraus muss der Schluss gezogen werden, dass die erhobenen Daten wirklich nicht mehr sind als ein synchroner Schnitt, und dass sich auch umfangreiche maschinenlesbare Lexikonsysteme sehr dynamisch entwickeln.

*

Für die 3 kleineren untersuchten Systeme gilt das Gesagte noch vermehrt, weil in kleineren Datenbeständen Reorganisation und Zuwachs leichter zu realisieren sind. Die Belege für diese These finden sich ebenfalls in den Fragebogen und im Begleitmaterial:

- HAM:

Im Fragebogen gibt es keine derartige Formulierung, dafür aber umso mehr in dem beigelegten Arbeitsbericht (Hoepfner83b), wo die in HAM-ANS verwendeten heterogenen lexikalischen Wissensquellen beschrieben sind:

a) Wortlexikon:

"Zur Zeit werden acht verschiedene Eigenschaften verwendet, deren Werte während der lexikalischen Analyse oder bei der Generierung flektierter Wortformen benötigt werden:" (p.1). (Diese Eigenschaften sind: Wortart, Stamm, Synonym, Flexion, Kasusrahmen, Genus, Deklination, Konjugation).

"Das Wortlexikon in HAM-ANS umfasst zur Zeit ca. 1250 Einträge." (p.1)

b) Syntagmenlexikon:

"Das Syntagmenlexikon enthält zur Zeit knapp 100 verschiedene Transformationen" (p.2) . (Diese Transformationen sind: Ersetzen, Komprimieren und Erwidern {Reaktion auf Benutzeräußerung}).

c) Kasusrahmenlexikon:

"Zur Zeit enthält das Kasusrahmenlexikon erst 8 Konstruktoren für verschiedene Ausprägungen von Kasusrahmen." (p.2/3)

d) Stammformenlexikon und Flexionsinformation:

"Die Flexionsanalyse und die Generierung flektierter Formen basieren zur Zeit noch auf einer unterschiedlichen Klassifikation von Grundformen, die sich jedoch ineinander überführen ließen, sodass in Zukunft beide Aufgaben durch den Zugriff auf eine gemeinsame Wissensquelle möglich sind." (p.4)

(Das Stammformen-Lexikon enthielt im Januar 1983 insgesamt 82 Einträge) (p.3).

e) Generell ist hier noch anzumerken, dass die gesamte "Konzeption der lexikalischen Analyse, wie sie für das System HAM-RPM [Vorläufer von HAM-ANS, H.H] entwickelt wurde", in HAM-ANS nicht übernommen wurde,

"sondern eine neue Kontrollstruktur entwickelt, die die Flexionsanalyse in die Syntagmenerkennung einbettet. Dies war auch aus dem Grunde notwendig, weil in HAM-RPM weitgehend mit einem Vollformenlexikon gearbeitet wurde." (p.6)

- KIT:

In M. Eimerachers Begleitbrief zu den 4 Arbeitsberichten, welche die Antwort auf den Fragebogen darstellen, finden sich viele zeitlichen Einschränkungen:

a) "in der Projektgruppe KIT werden zur Zeit zwei Parserverfahren verwendet".

b) "Für den GPSG-Ansatz gibt es bisher noch keine morphologisch-lexikalische Komponente".

c) "Ausserdem verzichten wir noch weitgehend auf platzsparende Kodierungen im Lexikon".

d) "Die lexikalische Komponente ist inzwischen in einer Reihe von Punkten weiterentwickelt worden, die jedoch nichts am prinzipiellen Aufbau ändern. Z.B. sind die Wortstämme im Lexikon in Buchstabenlisten zerlegt, so dass eine passende Wortform beim Lexikonzugriff (als Nebenprodukt) in Stamm und Endung zerfällt."

e) "Ausserdem ist das Verfahren auf Affixe und Suffixe so erweitert, dass beliebige Komposita verarbeitet werden können."

f) "Zahl der Stammformen zur Zeit: einige hundert."

- EVAR:

Die Hinweise auf zeitliche Einschränkungen lassen sich dem Fragebogen an folgenden Stellen entnehmen:

a) Frage Nr. 6 (File oder Datenbank?)

"momentan fileorientiert".

b) Frage Nr. 10 (Programmiersprache?)
"zukünftig in Lisp".

c) Frage Nr. 11 (Zugriffsalgorithmen?)
"noch unklar".

d) Frage Nr. 12 (Zugriffsschlüssel?)
"derzeit nur ein Schlüssel: Wortnummer".

e) Im Begleitbrief zum Fragebogen schreibt A. Brietzmann:
"Leider kann ich Ihren Fragebogen nicht in allen Einzelheiten beantworten, da unser Lexikon zur Zeit umstrukturiert werden soll (evtl. in eine Datenbank gebracht werden soll)".

*

Es wäre falsch, diese Sachlage für die gemachte Untersuchung negativ zu interpretieren.

Die häufig auftretenden Formulierungen zeitlicher Bedingtheiten deuten nicht auf eine schlechte empirische Datengrundlage hin, sondern bilden selbst empirische Indikatoren für den besonderen Charakter dieser Daten:

Solche sich verändernde Daten können als Prozessdaten oder als wissenschafts-historische Daten verstanden werden. Sie signalisieren Forschungsetappen. Es ist deshalb sinnvoll, den diachronischen Aspekt eingehender zu erörtern. Am ehesten kommt dafür die Auswahl eines relevanten Beispiels aus den 6 untersuchten Projekten in Frage, welches dann stellvertretend für die anderen diachron dargestellt werden kann.

Dieses Beispiel bildet das Lexikon EVAR, denn im Fragebogen und im Begleitmaterial finden sich starke Hinweise auf grundsätzliche Umstrukturierungen des Lexikonsystems, die gerade zur Zeit der Umfrage im Gang waren. Sehr deutliche Hinweise in dieser Richtung sind zwar auch in den erhaltenen Unterlagen zum System HAM-ANS festzustellen, aber diese beziehen sich teilweise auf eine frühere Umstrukturierung, die mit dem Übergang HAM-RPM --> HAM-ANS zusammenhängt.

Das Lexikon EVAR kommt auch deshalb noch eher für eine separate, diachrone Analyse in Frage, weil die dort zur Zeit der Umfrage stattfindende Umstrukturierung einen Trend repräsentiert, der auch bei grösseren Systemen aktuell war, nämlich den Übergang File --> Datenbank. (ALEXSYS --> SESAM).

Die Auswahl ist nicht im Sinne einer Wertung zu verstehen, sondern bildet lediglich eine Methode, die synchrone empirische Untersuchung auch diachronisch zu verankern.

3.3.2 Diachronische Aspekte.

Die diachronische Entwicklung der Lexikonkomponente kann anhand von Publikationen und Studienarbeiten in folgenden Stufen beobachtet werden:

1. Design im Rahmen der Systemarchitektur.
2. Realisierung einer Lexikonkomponente.
3. Revision und Ausbau der semantischen Komponente des Lexikons.
4. Überführung des Lexikons in eine Datenbank.
5. Präzisierung*in der Wörterbuchsyntax.
6. Pragmatik

Diese Stufenteilung basiert auf folgender Literatur:

- * H.-W. Hein, A System Understanding Continuous German Speech.
(Hein81)
- * R. Müller-Meernach, Datenbasis für den Lexikonexperten eines
Spracherkennungssystems.
(Müller-Meernach81)
- * U. Ehrlich, Semantische Klassifikation und Valenzbeschreibung
deutscher Adjektive und Substantive für das Lexikon der
automatischen Sprachverarbeitung.
(Ehrlich83)
- * A. Brietzmann, Semantische und pragmatische Analyse im Erlanger
Spracherkennungsprojekt.
(Brietzmann84a)
- * H. Endres, Ein Datenverwaltungssystem für ein Lexikon zur
automatischen Spracherkennung und Sprachgenerierung.
(Endres86)
- * U. Ehrlich, Ein Lexikon für das natürlichsprachliche Dialogsystem
EVAR.
(Ehrlich86)

1. Design des Lexikons im Rahmen der Systemarchitektur.

Grundsätzlich stellt das Lexikon eine Komponente im modularen Spracherkennungssystem dar, welche den folgenden Moduln (auch Experten genannt) lexikalische Wissensunterstützung gibt:

Akustik-Phonetik,
Phonologie-Prosodie,
Worterkennung,
Syntax,
Semantik,
Pragmatik und
Dialog

Dabei wird jeweils über Präprozessoren die für den entsprechenden Experten benötigte Lexikoninformation aus dem Nullexikon bezogen. Auf den Kontext dieser Architektur wurde in (2.2.3) schon eingegangen, sodass wir uns jetzt der Beschreibung der ersten Realisierung dieses Konzepts zuwenden können.

2. Realisierung der Lexikonkomponente.

Diese Komponente wurde 1980/81 im Rahmen der Studienarbeit "Datenbasis für den Lexikonexperten eines Spracherkennungssystems"

(Müller-Meernach81) realisiert. Dabei wurden u.a. die Gebiete

- Wortschatzbestimmung (a)
- Aufbau und Inhalt des Eintrags (b)
- Syntaktische Komponente (c)
- Semantische Komponente (d)
- Datenerfassung und Schnittstellen (e)

bearbeitet.

a) Wortschatzbestimmung

Die ursprüngliche Datenbasis war ein Vollformenlexikon mit einem

- Minimalwortschatz von 52 Struktur- und Funktionswörtern, 16 Substantiven 7 Adjektiven und 24 Verben (=96 Wörter), einem
- Fachwortschatz 'Intercity' aus 125 Wörtern wie "Platzkarte", "verreisen" u.a. und einem
- Ergänzungswortsschatz, der "die Zahlwörter enthält, durch deren Komposition sich der natürliche Zahlenraum von 0 bis 999'999 darstellen lässt. Dazu kommen die Wochentage, eine Liste der Monate, sowie einige Städtenamen."
(Müller-Meernach81, p.27)

"Alle drei Wortschätze enthalten zusammen 267 Einheiten"
(Müller-Meernach81, p.27)

Im Hinblick auf die Erweiterung dieses Bestandes werden folgende Methoden vorgeschlagen:

"Ausgehend von einem Rangwörterbuch [...] sequentiell die häufigsten Wörter wählen, jeweils beim ersten Auftreten einer flektierten Form die dazugehörige Grundform eintragen (mit allen möglichen Flexionen)"
(Müller-Meernach81, p.23)

"Wegen ihrer relativen Häufigkeit im Normaltext erst die Strukturwörter eintragen"
(Müller-Meernach81, p.23)

"Insbesondere für Verben berücksichtigen, ob sie im Valenzwörterbuch [Helbig69] enthalten sind. Dies führt zu einer Beschleunigung der semantischen Klassifikation"
(Müller-Meernach81, p.23)

b) Aufbau und Inhalt des Eintrags

Wie aus der letzten Bemerkung hervorgeht, enthalten die Lexikoneinträge auch semantische Information, und es ist auch gesagt, woher sie teilweise stammt. Wie der Eintrag gegliedert ist, kann aus der Syntaxbeschreibung abgeleitet werden, die in Teil 1, Kapitel (1.1.1(f)) wiedergegeben wurde. Allerdings ist in der Studienarbeit von 1981 der Aufbau des Eintrags noch nicht in dieser strengen Formalisierung gegeben, wie in der Syntaxbeschreibung von U. Ehrlich aus dem Jahre 1986.

c) Syntaktische Komponente

Diese Komponente wurde nach einer bereits vorgegebenen Schnittstellenbeschreibung zum Syntax-Modul erstellt.

"Die Schnittstelle von Lexikon und Syntaxexperte definiert sich durch die verwendete Grammatik. Die dort benötigten Kategorien werden auch im Lexikoneintrag aufgeführt. Allerdings besteht hier keine eindeutige Beziehung, der Lexikoneintrag ist etwas umfangreicher als die Schnittstelle erfordern würde."
(Müller-Meernach81)

Die Basis des syntaktischen Eintrags ist die Kategorie WORTART.

"Eine Liste der Wortarten sowie der dazugehörigen Wertzuweisungen findet sich im Anhang: "Manual für das Datenerfassungsprogramm". [...] Eine Erläuterung der in dieser Liste verwendeten Kategorien und Bezeichnungen findet sich in: (PLIDIS - Dokumentation Institut für deutsche Sprache Mannheim, 1980)"
(Müller-Meernach81, p.39)

d) Semantische Komponente

Für die semantische Komponente existierte damals noch keine Schnittstellenbeschreibung. Wie bereits unter

a) Wortschatzbestimmung
erwähnt, wurde ein Teil der semantischen Information aus (Helbig69) bezogen, wurde Dreistufenmodell wurde aber erweitert und auch die semantischen Merkmale sowie die Ordnung dieser Merkmale wurden modifiziert. Trotzdem gilt:

"Grundlage des Eintrages [für Verben, H.H.] ist die Verbvalenz."
(Müller-Meernach81, p.61)

Aber:

"Ausgehend davon wird versucht, die möglichen obligatorischen und fakultativen Aktanten als Realisierung von Tiefenkasus zu identifizieren."
(Müller-Meernach81, p.61)

Wie diese Kombination von Valenzmodell und Kasustheorie (nach Fillmore, Some Problems of Case Grammar, 1971) konkret aussieht, zeigt der in Figur 11 vorliegender Arbeit abgebildete Ausschnitt eines Semantik-Eintrags des Lexikonsystems EVAR.

Komplementär zu diesem den Verbeintrag betreffenden semantischen Ansatz wurde eine Klassifikation der Nomen vorgenommen. Über die Herkunft dieser Klassifikation schreibt Müller:

"Die Merkmale für die Nomen haben ihren Ursprung in einer Liste, wie sie in [D.Koch, W.Heicking, Syntactic-Semantic Analysis of German Sentences, in: L.Bolc (hg), Representation of Natural Language, München 1980, S. 207-270] vorgestellt wird, wurden jedoch noch ergänzt durch Merkmale, wie sie in einem Verfahren zur Substantivtypisierung des Linguistic Research System der University of Texas entwickelt wurden".
(Müller-Meernach81, p.66)

Neben den Verben und den Substantiven wurden auch die Präpositionen und die Subjunktionen semantisch klassifiziert. Alle semantischen Merkmale der verschiedenen Wortarten wurden unterschiedlich baumartig strukturiert, wobei auf Kompatibilität der Valenz- bzw. Kasusforderungen zu den Aktanten-, bzw. Rollenrealisierungen geachtet wurde.

e) Datenerfassung, Schnittstellen

Die wesentlichen Eigenschaften des Datenerfassungsprogramms für das Lexikon werden in (Müller-Meernach81) auf p.79 wie folgt angegeben:

"Das Programm führt den Benutzer durch den Lexikoneintrag: Sämtliche Kategorien werden vom Programm aufgerufen, der Benutzer muss noch die Wertzuweisungen eingeben.

Die Eingabe der Wertzuweisungen wiederum wird beschleunigt durch eine Vielzahl dabei verwendeter Abkürzungen.

Das Programm ist in der Lage, Eingaben des Benutzers auf Richtigkeit zu prüfen. Dies beschränkt sich jedoch natürlicherweise darauf, ob eine Wertzuweisung an die betreffende Kategorie grundsätzlich MÖGLICH ist. Ob sie im aktuellen Fall auch RICHTIG ist, muss der Benutzer selbst prüfen.

Das Programm übernimmt die optische Gestaltung des Lexikoneintrages.

Das Programm ist NICHT dazu geschrieben, um Änderungen in dem Lexikon vorzunehmen, dies soll mit einem Editor geschehen.

Das Programm wurde während der Erstellung eines 267 Einträge umfassenden Lexikons getestet."

Auf die Schnittstellen wurde bereits eingegangen. Sie sind im Zusammenhang mit der Konzeption "Nulllexikon" zu verstehen, welche bereits im Kapitel (2.2.3) erläutert wurde.

3. Revision und Ausbau der semantischen Komponente.

Die Erweiterungen, welche an der Lexikonkomponente vorgenommen wurden, sind in (Ehrlich83) beschrieben. Die Modifikationen beziehen sich auf vier Teilgebiete:

- Teilweise Umstrukturierung und Erweiterung des semantischen Klassifikationssystems.
- Übertragung des Valenzansatzes vom Verb auf Substantive und Adjektive.
- Erweiterung des (Tiefen-) Kasusystems.
- Erweiterung des interaktiven Erfassungsprogramms von (Müller-Meernach81) in bezug auf die semantischen Einträge von Adjektiven, Adverbien und Substantiven.

a) Klassifikation

Die grundsätzlichen Erweiterungen des Klassifikationssystems betreffen die Merkmalsordnungen der Adjektive und der Adverbien. Umstrukturierungen wurden auf der Basis eines früheren Systems für die Ordnungen der Substantive, Präpositionen, Konjunktionen und Subjunktionen vorgenommen:

"Dabei konnte ich mich bei Substantiven und Präpositionen auf ein schon vorhandenes System stützen [Müller-Meernach81], für Adverbien und Adjektive habe ich neue Klassifikationssysteme entwickelt, wobei auf Kompatibilität sowohl mit Semantikeinträgen bei Substantiven und Präpositionen, als besonders auch mit den Valenz-Einträgen des Verbs [Müller-Meernach81] und der Nomina zu achten war."

(Ehrlich83,p.7)

Anschliessend bemerkt U. Ehrlich:

"Die Klassifikation der Verben steht noch aus" (p.7)

Die semantische Klassifikation der Verben muss klar von der Beschreibung der Selektionsbeschränkungen der Aktanten bzw. der Kasusrollen der Verben unterschieden werden, denn die Klassifikation der Verben betrifft diese selbst und nicht deren Aktanten. Als Ansatz für die Verbklassifikation findet sich in (Ehrlich83) folgender Hinweis:

"Einige der abstrakten Substantivklassen könnte man auch auf Verben übertragen (ACT, PRO, RES, STT)."

(Ehrlich83,p.29)

Dabei bedeuten die Abkürzungen ACT = zweckgerichtetes Handeln, PRO = Process, RES = Result und STT = State. Als Begründung für diese Übertragbarkeit schreibt U. Ehrlich:

"Denn viele Verben übergeben ihren semantischen Inhalt bei der Substantivierung mit an das Nomen. z.B.

Ich fahre nach Hamburg (ACT)
Die Fahrt nach Hamburg (ACT)"

(Ehrlich83,p.29)

b) Übertragung des Valenzansatzes vom Verb auf Substantive und Adjektive.

So wie beim Valenzeintrag der Verben auf das 3-Stufen-Modell von Helbig/Schenkel zurückgegriffen werden konnte, liessen sich für die Realisierung der Valenzeinträge für Substantive und Adjektive die lexikographischen und linguistischen Vorarbeiten von K.-E. Sommerfeldt und H. Schreiber benutzen. (Bibliographiert als (Sommerfeldt83a) und (Sommerfeldt83b)). Zusätzlich wird in (Ehrlich83,p.40) auf W. Teubert, (Valenz des Substantivs - Attributive Ergänzungen und Angaben, Düsseldorf, 1979), hingewiesen.

Der hier verwendete Valenzbegriff wird ausdrücklich als "Valenz im weiteren Sinne" bezeichnet und wie folgt benutzt:

"Die Fähigkeit, Leerstellen zu eröffnen, besitzen nicht nur Verben, sondern auch andere Wortarten. Dies nennt man Valenz im weiteren Sinn. Einige Linguisten vertreten sogar die Ansicht, dass man theoretisch Wörtern jeder Wortart Valenz zusprechen kann. Im Rahmen dieser Arbeit wird die Valenz von Substantiven und Adjektiven besprochen. Dies sind die einzigen Wortarten, mit deren Valenz sich schon mehrere Linguisten ausführlicher beschäftigt haben (Sommerfeldt, Schreiber, Teubert). Mit einigen notwendigen

Änderungen kann man das Konzept der Verbvalenz auch auf die Valenz von Substantiv und Adjektiv übertragen."
(Ehrlich83, p.40)

c) Erweiterung des (Tiefen-) Kasussystems.

In bezug auf das Kasussystem von Ch. Fillmore wurde eine ähnlich weitreichende Erweiterung vorgenommen, wie es die in (b) beschriebene Übertragung des verbalen Valenzmodells auf die Wortarten Substantiv und Adjektiv darstellt.

"Fillmore's Kasussystem ist nur für Verben entwickelt worden. Kasusrahmen für Substantive können alle diese Kasus enthalten (man kann sehr viele Verben substantivieren, ohne dass sich Valenz und Kasusrahmen ändern), zusätzlich kommen aber noch Kasus hinzu, die es im verbalen Bereich nicht gibt. Im folgenden möchte ich erst auf die substantivischen Tiefenkasus, die identisch mit den verbalen sind, und ihre morpho-syntaktischen Entsprechungen eingehen, dann werde ich dieses System für die Substantive erweitern. Meine Beispiele dazu stammen überwiegend aus 'Valenz des Substantivs (Sommerfeldt83a)."
(Ehrlich83, p.52)

In (Ehrlich86) ist die Erweiterung des Kasussystems wie folgt resümiert:

"Für das Lexikon EVAR wurde das von Fillmore vorgestellte Kasussystem erheblich erweitert, vor allem für von Nomina abhängige Aktanten. Insgesamt besteht unser System jetzt aus 28 verschiedenen Tiefenkasus, die sich wieder in drei (nicht durchschnittsfreie Gruppen) einteilen lassen:

- Kasusrolle von Angaben
(das sind Satzglieder, die von allen Wörtern einer bestimmten Wortart abhängen können, unabhängig von der jeweiligen Wortbedeutung, z. B. lassen sich Konstituenten mit dem Tiefenkasus TIME mit allen Verben kombinieren, wie in 'ich arbeite heute' oder 'Peter fliegt erst morgen'.
- Kasusrollen von Aktanten, die von Nomina oder Adverbien abhängen. Dazu gehören z.B. EXPLICATION (die Gegend VON HANNOVER) und MEASURE (eine Höhe VON DREI METERN). Diese Gruppe schliesst auch alle Kasus der dritten Gruppe mit ein.
- Kasusrollen im verbalen Bereich.
Hier handelt es sich im wesentlichen um die von Fillmore vorgestellten Kasus (AGENT, OBJECT, EXPERIENCER, INSTRUMENT, SOURCE, GOAL, PATH, LOCATION und GOAL)".
(Ehrlich86, p.15)

Eine schöne graphische Darstellung des Kasusrahmens von 'fahren' ist in (Brietzmann84a), p.86 zu finden:

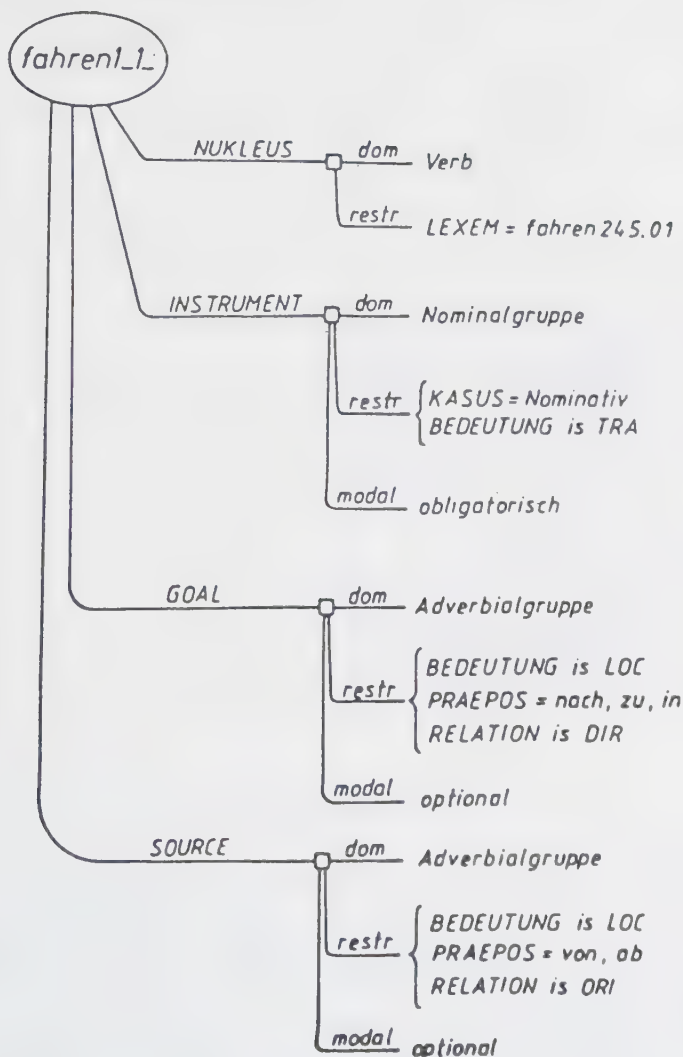


Figure 29. 'Kasusrahmen (EVAR)': (Brietzmann84a)

Die zusätzlichen Angaben zu den einzelnen Kasusrollenbezeichnungen betreffen

- syntaktische Realisierungsmöglichkeiten,
- Selektionsrestriktionen und
- die Angabe fakultativ versus obligatorisch.

Im Lexikon ist der abgebildete Kasusrahmen wie folgt formalisiert:

```
fahren (245)
  BEDEUTUNG 1: (von Fahrz.) sich fortbewegen
  OBLIGATORISCH 1
    KASUSFUNKTION: Nominalgruppe (KASUS: Nominativ)
    TIEFENKASUS: Instrument
    SELEKTION: Tra
  FAKULTATIV 1
    KASUSFUNKTION: Adverbialgruppe
                      (PRAEPOSITION: nach, zu, in;
                      SELEKTION: Dir )
    TIEFENKASUS: Goal
    SELEKTION: Loc
  FAKULTATIV 2
    KASUSFUNKTION: Adverbialgruppe
                      (PRAEPOSITION: von, ab;
                      SELEKTION: Dir )
    TIEFENKASUS: Source
    SELEKTION: Loc
```

(Brietzmann84a, p.87/88)

(Vergl. diesen Eintrag mit der weiter unten folgenden Form des Eintrags 'fahren')

Das in den Lexikoneinträgen formalisiert dargestellte lexikalisch-semantische Wissen wird in reformatierter Form dem Semantikmodul weitergereicht:

"Durch den Lexikon-Präprozessor wird das lexikalische semantische Wissen in mnemotechnische Abkürzungen codiert. Diese werden im Semantik-Modul ebenfalls benutzt und sind dort als globale Konstanten vereinbart. Die Schnittstelle wird festgelegt durch die Datei DLEXCOD, auf der der Lexikon-Präprozessor arbeitet." (Brietzmann84, p.121).

(Zum Prinzip des Lexikon-Präprozessors vergl. Kapitel 2.2.3)

Die Struktur des Semantikmoduls zeigt Figur 30

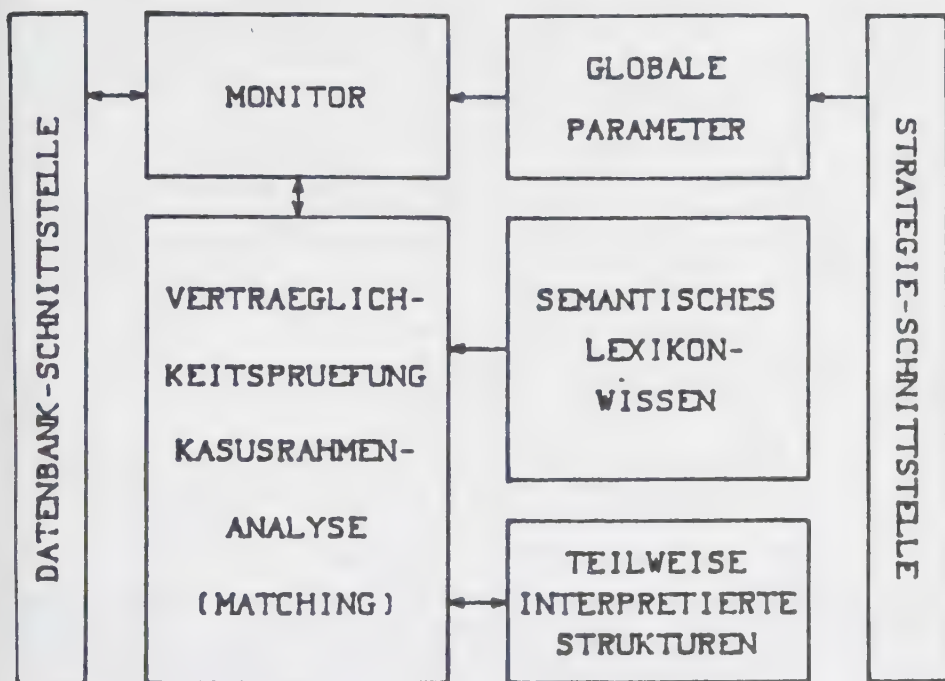


Figure 30. 'Semantik-Modul (EVAR)': In: (Brietzmann84a,p.102)

Der Semantikmodul weist ausser der erwähnten Schnittstelle DLEXCOD weitere Schnittstellen auf:

Die logischen Schnittstellen des Semantikmoduls sind teilweise über die Hypothesenbank (Datenbankschnittstelle) realisiert. Dies betrifft in erster Linie die Schnittstelle zum Syntax-Modul und die Schnittstelle zum Pragmatik-Modul. Die zentrale Aufgabe des Semantik-Moduls besteht darin, die vom Syntax-Modul gelieferten syntaktischen Baumstruktur-Hypothesen in textuell repräsentierte Kasusrahmen-Hypothesen zu transformieren. [**10] Diese Semantik-Hypothesen werden an den Pragmatik-Modul weitergereicht. Da die Hypothesenbank in verschiedene Abstraktionsebenen geschichtet ist (vergl. Kapitel 2.2.3), sind auch die Zugriffsfunktionen nach Ebenen geschichtet (Brietzmann84a,p.127)

(Auf die Schnittstelle zum Strategiemodul soll hier nicht näher eingegangen werden. Beschreibung in Brietzmann84a,p.125ff).

- d) Erweiterung des interaktiven Erfassungsprogramms von (Müller-Meernach81) in bezug auf die semantischen Einträge von

"Um den Eintrag der Wörter in die linguistische Datenbasis 'ALEXIKO' [=EVAR-Lexikon, H.H.] des Spracherkennungsprojekts zu erleichtern, habe ich das Datenerfassungsprogramm 'ISYS86' [(Müller-Meernach81), H.H.] um den semantischen Teil erweitert. Das Programm arbeitet im Dialogbetrieb. Die Programmiersprache ist Simula, die Systemhardware eine Cyber 173"

Das Programm erfragt alle Merkmale, die für eine Wortart bzw. eine bestimmte Funktion erwartet werden und überprüft, ob der Eintrag eine mögliche Wertzuweisung zum jeweiligen Merkmal ist. Wird eine Eingabe nicht akzeptiert, muss nur das zurückgewiesene Merkmal neu eingegeben werden, auch wenn vorher mehrere Wertzuweisungen auf einmal erfolgt waren.

Das Programm dient nur zur Aufnahme neuer Wörter in das Lexikon, Änderungen erfolgen mit Hilfe eines Text-Editors.

Für den semantischen Eintrag von Adjektiven und Adverbien wurde ein Zusatz zur Bedienungsanleitung verfasst."

(Ehrlich83, p.99)

5. Überführung des Lexikons in eine Datenbank.

"Das Lexikon lag bisher in Form von 15 Textdateien, ausgerichtet auf gute Lesbarkeit, vor. Es war zwar möglich, neue Wörter interaktiv einzugeben (Müller-Meernach81), erweitert durch (Ehrlich83), allerdings waren Suchmöglichkeiten höchstens über den Editor vorhanden. Eine Zusammenstellung von Teillexika nach bestimmten Kriterien hätte sich in einer Suche mittels Editor sowie anschließendem Kopieren von Teildateien erschöpft. Eine gezielte, linguistisch motivierte Arbeit war somit nicht möglich". (Endres86, p.6) M.a.W bedeutet dies, dass sich die Lexikon-Programme von (Müller-Meernach81) und die Erweiterungen durch (Ehrlich83) nicht in jeder Hinsicht als praktisch erwiesen und deshalb in einen neuen Ansatz -eine Datenbank- gefasst werden sollten.

Auf die vielen Funktionen dieser Datenbank (adaptierte Version von HBASE) wurde schon in (1.1.2(e)) eingegangen.

Durch die Vielfalt der Änderungs- und Auskunftsfunktionen wurde die Flexibilität erhöht:

"Einerseits ist es möglich, in das Lexikon vollständige Worteinträge oder nur Teile, z.B. syntaktische oder semantische Wörter neu einzubringen. Andererseits ist es aber auch erlaubt, Merkmale zu entfernen oder durch Spezifikation hinzuzufügen, so dass in der Zukunft beispielsweise auch Informationen über den Gebrauch von Wörtern in einem konkreten Anwendungsbereich ins Lexikon aufgenommen werden können". (Endres86, p.92)

Gerade dieser letzte Punkt ist im Rahmen der hier versuchten diachronischen Beschreibung der Entwicklung der Lexikonkomponente des Systems EVAR sehr interessant, weil hier angedeutet ist, dass neben den schon vorhandenen lexikalischen Informationseinheiten

- graphematische,
- phonologisch-phonetische,
- morpho-syntaktische und
- semantische Information

neu nun auch pragmatische Information ins Lexikon aufgenommen werden soll. In (Endres86) wird das zwar noch als Zukunftsperspektive bezeichnet, aber in der Zwischenzeit wurde ein derartiger Lexikonteil realisiert:

Marion Mast, Ein Pragmatik-Lexikon für den Anwendungsbereich "Intercity-Auskunft", Studienarbeit, Erlangen/Nürnberg 1986.

Diese Arbeit konnte von mir leider nicht mehr in die Untersuchung miteinbezogen werden, aber aus Gründen der Vervollständigung des diachronischen Überblicks musste sie hier unbedingt erwähnt werden.

*

Nun aber wieder zurück zum Thema 'Datenbank': Die Umstrukturierung des Lexikons in eine Datenbank hatte einen positiven Nebeneffekt bewirkt:

"Aufgrund der konsequenten Verwendung von Abkürzungen und der gedrängten internen Repräsentation der Einträge war es möglich, den Speicherbedarf bei der Auslagerung auf Dateien auf etwa die Hälfte zu reduzieren"
(Endres86, p. 94)

Die ebenfalls schon gezeigte Baumstruktur des Lexikoneintrages in EVAR (vergl. die Syntaxdarstellung in (1.1.1(f))) wurde in der internen Datenbankrepräsentation beibehalten:

"Bei Erwägung des Suchaufwandes, des Speicherbedarfs und des Problems der Zuordnung eines Attribut-Wert-Paares zu einem syntaktischen bzw. semantischen Wort zeigte sich letztendlich, dass die vorhandene hierarchische Gliederung eines Worteintrages auch für die interne Repräsentation günstig ist."
(Endres86, p. 92)

5. Präzisierung der Wörterbuchsyntax.

In (Ehrlich86) wird der neueste Stand der Lexikon-Syntax referiert: (d.h. derjenige von 1986, um es vorsichtiger zu formulieren)

"Die Form eines Lexikon-Eintrages soll [nochmals] am Beispiel des Wortes 'fahren' verdeutlicht werden:

```
[ WORT fahren
  WORTNR 245
  AUSSPRACHE ('.f.a.:.rrern.)

  SYNTAX-TEIL
    (I WORTARTNR 245.01
     WORTART Verb
     ...
     SEMANTIK-TEIL
```

```

([BEDEUTUNGSNR 245.01.01
BEDEUTUNG "der Zug faehrt von Hamburg
          ueber Nuernberg
          nach Muenchen"
...])

([BEDEUTUNGSNR 245.01.02
BEDEUTUNG "Peter faehrt mit dem Auto
          von Muenchen
          ueber Nuernberg
          nach Hamburg"
...])

([BEDEUTUNGSNR 245.01.03
BEDEUTUNG "Der Zugfuehrer faehrt
          den Zug"
...])

([BEDEUTUNGSNR 245.01.04
BEDEUTUNG "ein LKW-Fahrer faehrt
          Fisch von Hamburg
          ueber Nuernberg
          nach Muenchen"
...])

[WORTARTNR 245.02
WORTART Verb
...
REFLEXIVFAHIGKEIT
  [TYP nicht_echt_reflexiv
   KASUS Dativ]

SEMANTIK-TEIL
  ([BEDEUTUNGSNR 245.02.01
   BEDEUTUNG "der Junge faehrt ihr
             ueber das Haar"
...])])

[WORTARTNR 245.03
WORTART Verb
NORMALFORM ([WORT fahren
             WORTARTNR (245.01 245.02)])
           ([WORT abfahren
             WORTARTNR (270.01)
             VERBPRAEFIX
             [WORT ab
              WORTARTNR (411.04)])])
TEMPUS Praesens
PERSON (4 6)
MODUS indikativ ])]

```

(Ehrlich86, p.8/9)

Dieses Beispiel ist für EVAR repräsentativ, weil das Wort zum Fachwortschatz (Intercity) gehört und ein Verb ist. Die Primärquelle (Helbig69) wurde schon bei (Müller-Meernach81)

hervorgehoben. Zudem passt es gut zu der bereits dargestellten, ebenfalls auf "fahren" bezogenen Kasusrahmen-Darstellung. U. Ehrlich kommentiert das Beispiel wie folgt:

"Aus diesem Beispiel wird deutlich, dass in diesem Lexikon grammatische Wörter auch dann unterschieden werden, wenn sich ihre syntaktische Funktion nur in einem einzigen Merkmal unterscheiden (z.B. 245.01 und 245.02 unterscheiden sich durch das Merkmal der Reflexivität: 245.01 ist nicht reflexiv, 245.02 kann sowohl reflexiv als auch nicht reflexiv verwendet werden). Ein ebenso strenges Kriterium wird für die Abgrenzung einzelner Lexeme voneinander verwendet: einem grammatischen Wort werden immer dann verschiedene Lexeme zugeordnet, wenn es in mehreren Paradigmen auftreten kann (z.B. "ich fahre dir über das Haar"), oder wenn sie unterschiedliche Valenzrahmen aufbauen können (z.B. 245.01.01 - 245.01.02 - 245.01.03 - 245.01.04)" (Ehrlich86,p.8/9)

"Lexem" wird hier von U. Ehrlich im Sinne von "Bedeutungsvariante" verwendet. Es ist nicht mit dem Lemma zu verwechseln, welches in EVAR Normalform genannt wird und "als Repräsentant alle die Informationen trägt, die allen flektierten Formen eines Lexems gemeinsam sind." (Ehrlich86,p.6).

*

Zum Abschluss des diachronischen Überblicks soll darauf hingewiesen werden, wo das Lexikon, nach Ansicht einer beteiligten Verfasserin, noch verbessert werden könnte:

"In unserem Lexikon sind nur Wörter vorgesehen, die für sich allein die allgemein gebrauchten, 'normalen' Bedeutungen tragen und, mit den Einschränkungen der Valenz, Kombinationen mit anderen Wörtern eingehen können. Kann die Bedeutung nur einer Gruppe von Wörtern ('Syntagma') zugeschrieben werden, wobei die Grundbedeutung der einzelnen Wörter nicht erhalten bleibt, ergeben sich Probleme. Z.B.

Gebrauch machen von:

Ich mache Gebrauch von deinem Angebot.

Um diesen Inhalt zu realisieren, muss das Verb 'machen' mit dem Substantiv 'Gebrauch' zusammen verwendet werden. Es handelt sich nicht um eine obligatorische Valenzstelle, die durch alle Substantive einer semantischen Klasse bestzt werden kann, z.B.

(*) Benutzung machen von

[Der Stern deutet ungrammatischen Gebrauch an, H.H.]. Valenz muss hier der ganzen Wortgruppe, dem sogenannten Funktionsverbgefüge [(Teubert79)], bestehend aus Verb und Substantiv, in einigen Fällen zusätzlich aus Artikel und Präposition, zugesprochen werden. Z.B.

(mache Gebrauch)

(ich) (von deinem Angebot)

In unserem System ist dafür keine äquivalente Beschreibung vorgesehen.

Auch feste Wendungen (z.B. das Für und Wider) und idiomatische Ausdrücke (z.B. auf den Arm nehmen) können nicht im Lexikon dargestellt werden.

Langfristig wird daher für diese Fälle ein Syntagmen-Lexikon erforderlich sein."
(Ehrlich83,p.92)

3.3.3 Abschätzung des Erkenntniswerts der Untersuchung.

Der Nutzen der durchgeführten empirischen Untersuchung besteht hauptsächlich in der Zusammenstellung eines differenzierten Instruments zum Vergleich unterschiedlicher maschinenlesbarer Wörterbücher.

Durch die Auffächerung des Vergleichsrasters in 6 thematische Gruppen (3.1) und der synoptischen Präsentation der Antworten ist es möglich, den Vergleich zwischen den Wörterbüchern je nach Bedürfnis selektiv auf einen Aspekt zu fokussieren. Aus der zusammengetragenen Materialmasse kann deshalb für verschiedene spezifische Vergleiche eine Untermenge von relevanten Daten selektiert werden.

Exemplarisch wurde ein derartiger Vergleich in Bezug auf die in den 6 Lexika verwendeten semantischen Kategorien durchgeführt, wobei auch die zugeordneten Bezugssysteme und die zugrundeliegenden Semantikvorstellungen berücksichtigt wurden. Mit diesem Verfahren sollte u.a. demonstriert werden, dass es sinnvoll sein kann, Vergleiche auch an Systemen anzustellen, die, von ihren grundsätzlichen Konzeptionen her betrachtet, stark differieren.

Solange präzise festgehalten wird, auf welchen Aspekt der empirischen Datenmasse sich der Vergleich bezieht und auf welche Art die Daten in ihren ursprünglichen Bezugssystemen verankert sind, kann der Vergleich insofern sinnvoll sein, als er auf Lücken, Besonderheiten und den durchschnittlichen Stand der Entwicklung der Systeme aufmerksam macht.

Das Beispiel des Vergleichs semantischer Kategorien sollte aber auch dazu dienen, die Grenzen des Sinnvollen eines solchen Vergleichs zu zeigen, indem darauf verzichtet wird, "alles" vergleichen zu wollen, ohne sich auf ausgewählte Aspekte zu beschränken. Vom zusammengetragenen Material her betrachtet, hätte der Vergleich auch auf die Implementationsaspekte der 6 Systeme oder auf den Aspekt der morpho-syntaktischen Kategorien bezogen werden können. Dass dies nicht geschah, ist nicht auf das Datenmaterial zurückzuführen, sondern auf eine bewusste Selektionsentscheidung, die in (2.1.*) begründet wurde.

Durch diesen Entscheid unterscheidet sich der in dieser Arbeit durchgeführte Vergleich maschinenlesbarer Lexika von dem Forschungsbericht (Hess83) über maschinenlesbare deutsche Wörterbücher, wo ebenfalls vergleichend analysiert wird.

"Thema des Vorhabens war eine 'Vergleichende Analyse maschinenlesbarer deutscher Lexika'. Die Finanzierung des Projekts lag beim Bundesminister für Forschung und Technologie. Anlass, ein Vorhaben dieser Thematik in Angriff zu nehmen, war auf der einen Seite der Umstand, dass seit vielen Jahren Sprachanalyse- und -synthese-Verfahren entwickelt werden, die alle mit mehr oder weniger umfangreichen, doch unabhängig voneinander erstellten Lexika arbeiten. Diese auf den ersten Blick recht unterschiedlichen Lexika sollten hinsichtlich ihrer Vergleichbarkeit untersucht werden; dabei stand die Frage im Vordergrund, ob ein einheitliches integriertes Lexikon erstellt werden könnte, das für verschiedene Anwendungen geeignet wäre, so dass sich Entwicklungskosten im Bereich der Lexikonerstellung, die bisher mehrfach aufgewendet worden sind, einsparen liessen." (Hess83, p.1)

In (Hess83) geht es primär um die Abklärung der Möglichkeit der Erstellung eines integrierten Lexikons. Auf Fragen der lexikalischen Semantik wurde dabei nur am Rande eingegangen. So wie in der vorliegenden Arbeit bewusst die Semantik in den Vordergrund gestellt wurde, entschieden sich Hess et al. explizit für ein anderes Ziel, nämlich die Erstellung eines integrierten Lexikons. Dieses Ziel wurde auf der Stufe eines kumulierten Lexikons auch erreicht (vergl. Kapitel 2.2.5 Das Lexikonsystem ALEXSYS), wobei aber von den Verfassern auf den Seiten 215 ff des Berichts auf den Unterschied zwischen "kumuliertem" und "integriertem" Lexikon hingewiesen wird.

Die Motivation für ein alternatives Vorgehen, d.h. die Schwerpunktsetzung im Semantikbereich, ergibt sich u.a. auch aus der Tatsache, dass dieser Bereich in bezug auf maschinenlesbare Lexika bis anhin nicht systematisch-vergleichend behandelt wurde. Eine vergleichende Dokumentierung und Kommentierung semantischer Kategorien in maschinenlesbaren Lexika ist deshalb wünschbar. Darauf haben gerade auch Hess et al. aufmerksam gemacht:

"Motivation für dieses Projekt war dabei nicht nur das anwendungsbezogene Problem, aus früheren Arbeiten vorliegendes Material zusammen mit anderen Lexika für die Konstruktion neuartiger Lexikonsysteme verfügbar zu machen, es ging auch um die Problematik der angemessenen Darstellung von Informationen in der lexikalischen Mikrostruktur der maschinenlesbaren Lexika. Dass hierbei -angesichts der gegenwärtigen [1983, H.H.] Forschungslage- Probleme des semantischen Bereichs im Vordergrund hätten stehen müssen, versteht sich von selbst. Im vorliegenden Bericht, der im wesentlichen dokumentarischen Charakter hat, wird auf diese Probleme jedoch nur am Rande eingegangen." (Hess83, p.1/2)

Das bedeutet nun aber nicht, dass keine systematisch-vergleichenden semantischen Analysen von Wörterbüchern vorhanden sind. In der vorliegenden Arbeit wurde zur Interpretation eines bestimmten Typs von Semantikkonzeption (vergl. Kapitel 3.2.3) auf die vergleichende Arbeit von Neubauer - Die Struktur der Explikation in deutschen einsprachigen Wörterbüchern, eine vergleichende lexiko-semantische Analyse - zurückgegriffen (Neubauer80b). Neubauers Analyse legt zwar den Hauptakzent auf den semantischen Bereich, bezieht sich aber nicht auf maschinenlesbare Lexika. Der Vergleich semantischer Kategorien maschinenlesbarer Lexika wurde deshalb zum Schwerpunkt vorliegender Arbeit.

Die Kritik, welche an der Methode der durchgeführten Datenerhebung mittels Fragebogen im vorliegenden Fall am ehesten vorgebracht werden könnte, betrifft den besonderen Charakter der beobachteten Systeme: EDV-Systeme sind sehr dynamisch, und "aktuelle Daten" werden bald zu sog. "Steinzeit"-Relikten, wie sich viele EDV-Praktiker auszudrücken pflegen.

Dieser kritische Punkt wurde bei der durchgeführten Untersuchung durchaus im Auge behalten, indem empirisches Material auch nach dem Zeitpunkt der Datenerhebung weitergesammelt wurde, sodass es möglich war, den diachronischen Aspekt eines der untersuchten Lexikonsysteme zu beschreiben.

Obwohl der Untersuchungszeitraum durch Einbezug des diachronischen Aspekts ausgeweitet wurde, lässt sich eine im Vorwort zum Fragebogen (vergl. Anhang) geäußerte Vermutung nur negativ beantworten. Die zum Zeitpunkt der Abfassung des Bogens (1984) gemachte Annahme lautet:

"Eventuell lassen sich bei der Auswertung Ihrer Antworten gewisse Trends zu einer Standardisierung maschineller Lexika im deutschen Sprachbereich feststellen"

Aufgrund der eingegangenen Antworten und deren Vergleich lässt sich diese Hypothese aber klar verwerfen. Das gilt im einzelnen für folgende Lexikoncharakteristika, bei denen es prinzipiell denkbar wäre, dass gewisse Normen oder Übereinstimmungen zu beobachten wären:

1. Merkmale auf linguistisch-lexikographischer Ebene,
2. auf der Ebene der Datengrundlage der Lexikonsysteme,
3. auf der Ebene der eingesetzten informatischen Mittel,
4. in bezug auf definierte Schnittstellen und
5. Merkmale in bezug auf die Semantik-Konzeptionen.

Durch Verweise auf die in vorliegender Arbeit gesammelten und kommentierten Daten kann dieser negative Befund in bezug auf die 5 genannten Merkmale überprüft werden.

1. Auf linguistisch-lexikographischer Ebene lassen sich bei den 6 untersuchten Systemen für folgende Punkte wenig, bzw. keine Übereinstimmung feststellen:

a) Zahl und Qualität der Eintragskategorien. (Die Belege finden sich in Kapitel 3.1.1 Abschnitt [d], wo die Antworten auf die 4. Frage des Bogens vergleichend kommentiert werden. Die Informationsfelder pro Eintrag sind dort für jedes der 6 Systeme u.a. in Tabellen belegt.)

b) Wortklassenschemata. (Die Belege finden sich ebenfalls in Kapitel 3.3.1, jedoch in Abschnitt [c]). Die Divergenzen bei den Wortklassen sind umso bemerkenswerter, weil Wortklassen primär ein morpho-syntaktisches Phänomen darstellen und deshalb anzunehmen war, dass eine einheitlichere Spezifikation eher möglich wäre als

auf semantischer Ebene.

c) Morphologischer Typ. Die Unterschiede bezüglich der Gestalt des Lemmas sind in Kapitel 3.3.1, Abschnitt [b] belegt und zeigen das ganze Spektrum von Stammformen über Grundformen bis zu Vollformen sowie Mischtypen.

Auf die Nennung weiterer Unterschiede sei hier verzichtet, in Kapitel 3.3.1 sind aber weitere zu finden.

2. Die unterschiedlichen Datengrundlagen (lexikalische Quellen) sind in Kapitel 3.1.2 belegt.

3. Die in den 6 untersuchten Lexikonsystemen eingesetzten informatischen Mittel fächern sich sowohl im Bereich der Software als auch der Hardware weit auf (Kapitel 3.1.3 und 3.1.6).

4. Genormte Schnittstellen, die einen Austausch der Daten ermöglichen würden, wurden zum Zeitpunkt der Untersuchung nicht verwendet. (Kapitel 3.1.5)

5. Semantik-Konzeption: Obwohl man versucht sein könnte, gerade auf diesem Gebiet besonders grosse Unterschiede feststellen zu können, weil das Wissen auf semantischem Gebiet weniger gesichert ist als etwa auf morphologischer Ebene, ist das Gegenteil der Fall. In den Kapiteln 3.2.1 bis 3.2.3 ist belegt, dass sich die Semantik-Konzeptionen in 3 von 6 Systemen zum Zeitpunkt der Untersuchung erheblich gleichen. Auf diese Ausnahme in bezug auf den generell negativen Befund hinsichtlich Standards und Normen wird noch eingegangen.

*

Trotz dieses negativen Befundes ist anzunehmen, dass sich auf dem Gebiet der maschinenlesbaren Wörterbücher langfristig vermehrt Standards durchsetzen werden, weil auf diesem Gebiet das Entstehen einer eigentlichen "Sprachindustrie" beobachtet werden kann. Zumindest wurde diese Ansicht in einem Referat im Rahmen des von der DFG (Deutsche Forschungsgemeinschaft) organisierten speziellen Forums "Standardization in computerized lexicography" zum generellen Thema "Information Science and Practice" vertreten:

"There are concrete expectations with regard to data processing which must be seen in a context of applications which are focussed by a rapidly growing 'language' industry".
(Drewek86, p.8)

Nach einer vorsichtigen Interpretation ökonomischer Prognostikzahlen aus US-Quellen (Datamation 15, 1986, p.3) für den besagten Sektor, fährt R. Drewek fort:

"It is trivial to annotate that most of the systems which grow up inside research and development departments need lexical data more or less extensively. While dialog interfaces tend to be more domain specific, text editing, machine translation and content scanning show an increasing demand for lexical data. In each case, much of the effort will go into building up vocabularies and defining semantic relationships. Despite of different subject matters there must be some coordination and harmonization."

(Drewek86, p.8)

Als Träger eines möglichen Trends in Richtung Standardisierung und Normierung könnte sich u.U. das Verlagswesen erweisen:

"The 'classical' language industry, which has been producing dictionaries already when computers played no essential role, will get a key position if she succeeds in achieving an appropriate technological state and compiles just those data, which are useful in the office or business environment as well as in scientific and technical fields. Unfortunately these areas do not overlap with lexicographic work in the historical dimension and only partially with the colloquial language field. Nevertheless we can expect that the development of tools and techniques triggered from a more business oriented side will be helpful too in the scientific area." (Drewek86, p. 8/9)

*

In der vorliegenden Untersuchung maschinenlesbarer Lexika können Indikatoren für den Trend zu applikationsorientierter Computer-Linguistik ausgemacht werden, die sich vom Ansatz theoretischer CL unterscheiden.

Ein Beispiel für die ingenieurwissenschaftliche Behandlung eines linguistischen Problems stellt die Verwendung der Kasus Theorie in den Systemen HAM, KIT, EVAR dar. In Kapitel 3.2.3 wurde darauf hingewiesen, dass in diesen drei Systemen nicht besonders darauf geachtet wurde, eine originale Form der Kasusgrammatik im Sinne von Ch. Fillmore zu realisieren, sondern dass es eher darum ging, eine operationalisierbare Theorie zu implementieren, welche bei der Transformation syntaktischer in semantische und pragmatische Repräsentationen einsetzbar ist. Das könnte auch eine Erklärung für die oben erwähnte Ausnahme im Hinblick auf übereinstimmende Standardvorstellungen abgeben. Wie sich solche de facto Standards aus einem ingenieurwissenschaftlich verstandenen Ansatz in der Computer-Linguistik ergeben haben könnten, soll im folgenden mit drei Beispielen angedeutet werden, wobei aber auf eine ausführliche Erörterung der Vermutung verzichtet wird. Das erste Beispiel bezieht sich auf die Verwendung der Kasusgrammatik:

"Dieser Abschnitt verzichtet weitgehend auf die Erklärung der linguistischen Theoriebasis der Kasusgrammatik. Sie soll nur soweit erläutert werden, dass deutlich wird, was ein Lexikoneintrag, der Kasusrahmen verwendet, enthält, und wie er erstellt wird" (Müller-Meernach81, p.56)

Immerhin verweist Müller an dieser Stelle auf folgende Literatur zur Kasusgrammatik:

- W. Abraham, Kasus Theorie, Wiesbaden, 1971
- B. Bruce, Case Systems for Natural Language, in Artificial Intelligence, Vol. 6, 1975
- (Fillmore71)

Der ingenieurwissenschaftliche Ansatz wird vielleicht noch klarer, wenn ein zweites Beispiel betrachtet wird:

"Ausgehend von den Forderungen 1-4 aus Abschnitt 3.4.2 (Es handelt sich dabei um die Systemvorgaben 'Aufgaben des Semantikexperten', denen in der Lexikonkomponente Rechnung getragen werden soll:

1. Verträglichkeitstests, 2. Hypothesenbildung, 3. Rol-

lenidentifizierung, 4. Übergang Syntax --> Pragmatik H.H.1 möchte ich nun zwei linguistische Ansätze herausgreifen, die ausdrücklich das Verb als den Mittelpunkt der syntaktischen Struktur und des semantischen Gehalts eines Satzes darstellen. Es geht dabei zum einen um die Verbvalenz, eine Klassifikation auf dieser Grundlage ist eine syntaktisch motivierte [...]. Zum anderen handelt es sich um Kasusrahmen für Verben, ein Klassifikationsansatz, der semantisch motiviert ist [...]. Beide Verfahren sollen zu einem sinnvollen Beschreibungsmodell kombiniert werden.
(Müller-Meernach81, p50)

Neben diesen zwei Beispielen (Operationalisierung der Kasusgrammatik und modellkonforme Kombination von Kasusgrammatik und Valenztheorie [**11]) kann auf ein weiteres Beispiel hingewiesen werden, welches in Kapitel 3.3.2 erwähnt wurde. Es handelt sich um die Übertragung des Valenzansatzes vom Verb (nach G. Helbig, 1975) auf Substantive und Adjektive (nach E. Sommerfeldt und H. Schreiber). Diese drei Beispiele beziehen sich alle auf das Lexikon des Systems EVAR. Als Beleg für das ingenieurwissenschaftliche Denken mögen sie genügen, wobei damit nicht gesagt ist, dass in den Projekten KIT und HAM mit grundsätzlich anderen Methoden gearbeitet wird.

*

Zusammenfassend kann zum Schluss gesagt werden, dass die vorliegende vergleichende Untersuchung von 6 Systemen maschinenlesbarer Lexika u.a. folgende Themen behandelt:

- Semantikvorstellungen zum Zeitpunkt der Untersuchung,
- Gebiete, auf welchen Standards fehlen, aber denkbar wären und
- Ansätze ingenieurwissenschaftlicher Linguistik.

ANMERKUNGEN

Die Anmerkungen, wurden auf ein Minimum beschränkt, weil im Kontinuum des Textes auch gesagt werden kann, was sich in Anmerkungen ausdrücken liesse. Eine Ausnahme bilden die direkt in Zitate eingeflochtenen Erläuterungen, die teils vom Verfasser des Zitats, teils von mir gemacht werden.

ANMERKUNGEN ZU TEIL 1.

[**1]

Die DIN (Deutsche Industrienorm) 2333 ist beschrieben in:

- DIN Taschenbuch 153 "Publikation und Dokumentation 1, Normen für Verlage, Bibliotheken, Dokumentationsstellen, Archive", p.183. (Hg. DIN Deutsches Institut für Normung e. V.; Beuth-Verlag, Berlin, Köln).

Die Normen werden jeweils dem neusten Stand angepasst.

Die DIN 2333 heisst "Fachwörterbücher, Stufen der Ausarbeitung". Der im Zusammenhang mit "Abfolgeschema" entscheidende Begriff aus DIN 2333 heisst "Wortstelle". Er ist in der Norm so festgelegt:

"Eine Wortstelle (E entry, article oder item; F rubrique oder article) ist die kleinste selbständige Einheit in einem Fachwörterbuch, d.h. die Darstellung einer einzigen Zuordnung zwischen Begriff (E concept; F notion) und Benennung (E term; F terme). In einem systematischen Fachwörterbuch enthält eine Wortstelle - dann auch 'Begriffsstelle' genannt - alle Angaben über einen einzigen Begriff. Je nach dem Zweck des Fachwörterbuchs gibt sie entweder nur die dem Begriff zugeordnete (Vorzugs-)Benennung an, bei einem mehrsprachigen Fachwörterbuch auch die entsprechenden Benennungen in den anderen Sprachen, oder sie enthält zusätzlich noch weitere Aussagen über den Begriff, wie Ordnungsangaben, Definitionen, Synonyme (neben der Vorzugsbenennung noch zugelassene und/oder abgelehnte Benennungen), Bilder und andere Angaben." (p.184)

Im Zusammenhang mit der DIN 2333 sind noch drei weitere Normen zu erwähnen, die in einem direkten Zusammenhang mit 2333 stehen und ebenfalls im Normentaschenbuch 153 zu finden sind:

- DIN 2336 "Lexikographische Zeichen für manuell erstellte Fachwörterbücher" (p.193).

- DIN 2331 "Begriffssysteme und ihre Darstellung" (p.149).

- DIN 2330 "Begriffe und Benennungen, Allgemeine Grundsätze" (p.130).

[**2]

Die gemachte Unterteilung in DML und DDL ist nicht zwingend. Im Bereich relationaler Datenbanktechnik wäre SQL ein Gegenbeispiel, wo die Data Definition und die Data Manipulation in der gleichen Sprache ermöglicht wird.

[**2]

Dieser Syntaxbegriff ist an dieser Stelle weniger strikt gebraucht, als dies in der Informatik durch den Backus/Naur Formalismus geschieht. Es ist in diesem Zusammenhang an den in (Petöfi77) auf den Seiten p.263ff explizierten Gebrauch gedacht. "Das Lexikon als Komponente einer semiotischen Theorie von Sprachen" (p.287) wird dabei u.a. syntaktisch analysiert. Der dort verwendete Syntaxbegriff ist in folgenden Zusammenhang gestellt:

"Die Bezeichnung 'semiotische Theorie' im Titel weist darauf hin, dass das Lexikon als Teil einer Sprachtheorie verstanden wird, deren Ziel es ist, die Strukturen/Mitteilungen verschiedenster Größenordnungen einer natürlichen Sprache syntaktisch, semantisch und pragmatisch explizit zu beschreiben" (Petöfi77, p.287)

An anderen Stellen vorliegender Arbeit kann 'Syntax' eine andere Bedeutung haben als die eben erwähnte, z.B. in der in (1.1.1(f)) gezeigten Syntaxbeschreibung für das Lexikon EVAR.

[**3]

Diese Unterscheidung verschiedener Syntaxdefinitionen ist in der Datenbanktechnik jedoch nicht zwingend und hängt vom angewandten System ab. Als Gegenbeispiel zur Verwendung unterschiedlicher Syntaxen für DML und DDL kann die Datenbanksprache SQL (Structured Query Language) genannt werden.

[**4]

Es handelt sich um folgende Publikationen:

- (Quillian67), siehe Bibliographie.

- M.R. Quillian, Semantic Memory, In: M. Minsky (ed), Semantic Information Processing, Cambridge (Mass.), MIT, 1968, p.227-270.

- M.R. Quillian, The Teachable Language Comprehender. In: Communications of the ACM, 12, 1969, p.459-476.

[**5]

Die im folgenden aufgezählten Verfahren werden nicht nur beim Retrieval, sondern auch beim Indexieren und beim Abspeichern von Dokumenten angewandt. Diese bei U. Hahn genannten linguistischen Verfahren sind in der IR-Technik nicht die einzigen Methoden, welche bei Retrieval, Indexieren und Speicherung zur Anwendung kommen. Falls informationslinguistische morphologische, syntaktische oder semantische Verfahren eingesetzt werden, sind diese in weitergehende Modelle eingebettet, welche nicht auf linguistischen Konzeptionen beruhen. Vergl. dazu (Mresse84) p.24ff "Verwendungsarten von Thesauri" und p.29ff "Erkennung physisch ähnlicher Zeichenketten".

[**6]

An dieser Stelle merken W. Schulze und G. Heinze an:

"Aufgrund ihres kumulativen Charakters enthält die Wortdatenbank sowohl Grund- als auch Vollformen. Die Definition dieser Typen folgt herkömmlichen linguistischen Gesichtspunkten."
(p.42)

[**7]

"Einzelfile" bezieht sich auf die 4 grossen Wortartfiles der kumulativen Wortdatenbank:

- Nomen
- Verben
- Qualifier
- Funktionswörter

[**8]

An dieser Stelle merken W. Schulze und G. Heinze an:

"Die Lemmatisierung erfolgt aufgrund der Annahme von 'Performanz-', 'Kompetenz-' und 'Referenzlemmata'; vgl. bes. W. Schulze/ G. Willée, Noch eine Variation über das Thema 'Lemma'. In: Sprache und Datenverarbeitung 1/83, Tübingen, 1984." bibliographiert als (Schulze/Willée84).

[**9]

An dieser Stelle merken W. Schulze und G. Heinze an:

"Die Angabe der 'Normalform' (NF) (so MOLEX) oder paralleler Informationen in SADAW erfolgt nach einem vereinheitlichten Prinzip".

Dieses Prinzip ist die Unterscheidung von 'Performanz-', 'Kompetenz-' und 'Referenzlemmata' nach (Schulze/Willée84).

[**10]

In (Zehnder85) sind auf den Seiten 243/244 die einschlägigen Publikationen von E.F. Codd bibliographiert.

[**11]

Anomalien können im Zusammenhang mit Inkonsistenzen gesehen werden. Dabei können nach (Zehnder85,p166/67) modellinhärente von modellexternen Konsistenzbedingungen unterschieden werden:

"Modellinhärente Konsistenzbedingungen können mit Mitteln der verfügbaren Systemsprache direkt dargestellt werden; sie werden durch die zugehörigen Compiler und Laufzeitsysteme automatisch sichergestellt"
(Zehnder85, p.166)

"Modellexterne Konsistenzbedingungen werden in einer geeigneten allgemeinen Programmiersprache zur Darstellung der nicht modellinhärent formulierbaren Bedingungen ausformuliert"
(Zehnder85, p.167)

[**12]

P.P. Chen, Entity-Relationship Approach to Systems Analysis and Design, Amsterdam, 1980

[**13]

Für Datenbankankwendungen im Bereich des Information Retrieval trifft diese Feststellung nicht zu. Es sei deshalb auf das in Abschnitt g) des vorliegenden Kapitels erwähnte Modell von L. Rostek verwiesen.

[**14]

Auf dieses sehr anschauliche Beispiel hat mich freundlicherweise M. Mresse hingewiesen.

[**15]

Dieses Arbeitstreffen fand in Bonn statt und dauerte vom 27.1 bis zum 28.1 1983 (Hein83).

[**16]

Weitere Verfahren sind z.B.:

- G. Thurmair, Die Konzeption der morphologischen Analyse im Projekt CONDOR (Siemens). In: D. Krallmann (Hg.), Zur Morphologie des Deutschen, Hamburg 1980.

- H. Bergmann, Lemmatisierung in HAM-ANS. (Bergmann82)

[**17]

An dieser Stelle kann auf den Zusammenhang zwischen Sinnrelationen und Definitionen hingewiesen werden. Begriffslogische Relationen dienen sowohl der Realisierung von Verweisen zwischen Einträgen über die als Begriffe verstandenen Lemmata als auch zur internen begrifflichen Strukturierung von Definitionen. Im ersten Fall sind sie Teil der Verweisstruktur, im zweiten aber Teil der Mikrostruktur.

In der Lexikographie werden u.a. folgende Definitionstypen unterschieden:

"Typische Beispiele sind die folgenden, die allesamt dem Ullstein-Lexikon der deutschen Sprache (Frankfurt 1969) entnommen sind:

die synonymische Definition (kulant = grosszügig, umgänglich, zuvorkommend),
die antonymische Definition (massiv = nicht hohl),
die hyperonymische Definition (schwelen = glimmend, ohne eigentliche Flamme, langsam brennen),
die Negationsdefinition (nachgeben = einem Druck keinen Widerstand entgegensetzen),
die konjunktionale Definition (akut = plötzlich und heftig auftretend),
die metonymische Definition (Fuss = unterster Teil des Beines)."
(Hausmann85, p.373)

Es ist schwierig, alle diese verschiedenen Definitionsarten einheitlich zu beschreiben. Nicht ohne guten Grund lässt es E. Wüster in seiner "Einführung in die Allgemeine Terminologielehre und Terminologische Lexikographie" bei einer allgemein gehaltenen Definition der Definition bewenden:

"Eine Definition (Begriffsbestimmung) im weiteren Sinne ist die Beschreibung eines Begriffes durch bekannte Begriffe, und zwar meist mit Hilfe von Worten."

(Wüster79, p.29)

Die Art der Definition wird determiniert

a) durch die Art der Begriffsverknüpfung (Relation) zwischen dem Definiendum und dem Definiens und

b) durch die Art der Begriffsverknüpfung innerhalb des Definiens, falls dieses aus mehreren Begriffen besteht.

Die von Hausmann genannten oben zitierten Definitionstypen aus dem Ullstein Lexikon

- 'Synonymdefinition',
- 'Antonymdefinition' und
- 'Negationsdefinition'

sind bereits durch die Art der Relation zwischen Definiendum und Definiens determiniert (Fall (a)).

- Die 'hyperonymische' Definition,
- die 'konjunktionale' Definition und
- die 'metonymische' Definition

hingegen sind komplexer (Fall (b)). Hier ist die Definition mehr als die Angabe eines definitorischen Begriffs und der Sinnrelation zwischen dem Definiendum und diesem Begriff. Das Definiens besteht hier in Übereinstimmung mit der Definition von Wüster in einer "Beschreibung durch bekannte Begriffe" (Mehrzahl).

Es scheint deshalb sinnvoll zu sein, Definitionen in einfache und komplexe aufzuteilen.

Die einfachen Definitionen sind durch Angabe einer Sinnrelation und eines erklärenden Begriffs hinreichend bestimmt. Aber auch diese Art der Definition ist keineswegs trivial, wenn man die grosse Zahl von Sinnrelationen betrachtet. (Vergl. (Lutzeier85), p.106ff)

Die in (Wüster79) im Detail beschriebenen "klassischen" Definitionstypen 'Inhaltsdefinition' und 'Umfangsdefinition', sind prinzipiell komplexer als die eben genannten.

"Eine Inhaltsdefinition - oder Definition im engeren, klassischen Sinne - besteht in der Angabe des Begriffsinhalts. Sie geht von einem bekannten Oberbegriff aus und nennt die einschränkenden Merkmale, die den zu definierenden Begriff kennzeichnen und von den anderen Begriffen derselben Reihe unterscheiden." (Wüster79, p.30)

"Eine Umfangsdefinition besteht in der Aufzählung des Begriffsumfanges; d.h. aller Unterbegriffe, die im Begriffssystem auf derselben Stufe stehen. Bisweilen kommt auch die Aufzählung der Individuen in Frage.

Eine Umfangsdefinition kann nicht für alle Zeit erschöpfend sein, denn im Laufe der Zeit können weitere Unterbegriffe, z.B. neue Konstruktionen, entstehen, die unter dieselbe Inhaltsdefinition fallen.

Oft aber ist eine Umfangsdefinition leichter verständlich als eine Inhaltsdefinition. Man kann die Vorteile beider Definitionsarten vereinigen, wenn man eine Inhaltsdefinition für einen bestimmten Begriff durch eine Umfangsdefinition ergänzt, d.h.

veranschaulicht."
(Wüster79, p.31)

Das gemeinsame Merkmal komplexer Definitionen besteht darin, dass die definitorische Paraphrase in sich eine begriffslogische Binnenstruktur aufweist und diese komplexe Struktur als Ganzes mit dem Definiendum relationiert wird.

Sowohl bei der komplexen als auch bei der einfachen Definition kann entweder auf ein offenes Vokabular zurückgegriffen werden oder auf ein kontrolliertes. Auch können allfällig entstehende Explikationsketten durch den Aufbau eines Sets von kontrollierten Vokabeln vermieden (rückgeführt, abgebrochen) werden oder aber durch Verweise auf andere Lemmata im gleichen Wörterbuch explizit gemacht werden. (Verweisstruktur).

Weil Definitionen begriffslogisch strukturiert und untereinander relationiert sind, werden auch die Verweise vielfältiger als beim einfachsten Fall des Synonymthesaurus.

ANMERKUNGEN ZU TEIL 2.

[**1]

USL bedeutet User Speciality Language. Vergl. dazu

- N. Ott, M. Zoeppritz, USL - an Experimental Information System based on Natural Language, 1979. In: L. Bolc (hg), Natural Language Based Computer Systems, München 1980. (Bolc80).

- M. Zoeppritz, Syntax for German in the User Speciality Language System, Tübingen, 1984.

H. Lehmann selbst verweist auf:

- F. Guenther, H. Lehmann, Automatic Construction of Discourse Representation Structures. In Proceedings COLING'84, p.398.

[**2]

DRS bedeutet Diskursrepräsentationsstruktur. Es handelt sich dabei um eine Repräsentation, die von Hans Kamp vorgeschlagen wurde. Vergl. (Kamp81). H. Lehman zum Begriff DRS:

"Der zentrale Begriff der Diskursrepräsentationstheorie ist die Diskursrepräsentationsstruktur (DRS), die als Resultat einer satzweisen Übersetzung eines Textes nach sogenannten Diskursinformationsregeln erfolgt. DRSen können dann nach dem üblichen Vorgehen der Prädikatenlogik modelltheoretisch interpretiert werden.

Eine DRS ist ein Paar $\langle U, C \rangle$, wobei U eine Menge von Diskursreferenten und C eine Menge von Bedingungen ist. Die Bedingungen können atomar oder komplex sein. Atomare Bedingungen sind Literale wie $P(a,b)$ oder $a=b$, komplexe Bedingungen bestehen

aus einem Operator (Implikation, Disjunktion, Negation) und ein oder zwei DRSen." (Lehmann85, p.95)

[**3]
H. Lehmann verweist auf
- F. Guentthner, H. Lehmann, Rules for Pronominalization. Proc. 1st Conference and Inaugural Meeting of the European Chapter of the ACL, Pisa, 1983.

[**4]
"Zur Realisierung eines Expertensystems, das den gezeigten Dialog mit dem Benutzer zu führen vermag, haben wir folgende Komponenten entworfen:

1. Eine Dialogkomponente: führt den Dialog mit dem Benutzer. Sie leitet natürlichsprachliche Eingaben an den Natural Language Analyzer weiter, ist zuständig für die Erzeugung von Rückfragen und Erklärungen (gemäß dem Beweisbaum, der von der Deduktiven Komponente erstellt wurde).
 2. Der Natural Language Analyzer basiert auf dem USL-System, dessen Grammatik so erweitert wird, dass fortlaufende Texte behandelt werden können. Der Natural Language Analyzer erzeugt die zu einem Text gehörige Diskursrepräsentationsstruktur, löst kontextuelle Referenzen auf (wofür falls notwendig die Deduktive Komponente aufgerufen wird), und enthält ein Programm zur (halb-) automatischen Erweiterung eines Thesaurus von Bedeutungsregeln (Verweis auf: R. Wirth, Halbautomatische Erstellung eines bestehenden Thesaurus, Diplomarbeit, Heidelberg, 1984.1).
 3. Die zum Text gehörige DRS wie auch die sonstigen Regeln, die zur Behandlung einer Anfrage benötigt werden, werden vom Knowledge Manager verwaltet und bereitgestellt. Dieser wird zumeist vom Natural Language Analyzer und von der deduktiven Komponente aktiviert und ist auch zuständig für die Kommunikation mit der Datenbank, die die gesamte Wissensbasis enthält.
 4. Die Deduktive Komponente wird aktiviert, wenn Anfragen des Benutzers zu beantworten sind oder wenn kontextuelle Referenzen aufzulösen sind. Sie ist auf die Diskursrepräsentationstheorie zugeschnitten und basiert auf dem Tableauealkül (näheres siehe Guentthner / Lehmann / Schönfeld, A Theory for the Representation of Knowledge, eingereicht beim IBM Journal of Research and Development). Die Deduktive Komponente kann ihrerseits wieder die Dialogkomponente aktivieren, wenn Rückfragen an den Benutzer zu stellen sind."
- (Lehmann85, p.93)

[**5]
Ich möchte an dieser Stelle den im Vorwort ausgesprochenen Dank an diese Personen wiederholen. Sie haben mir auf kollegiale und liebenswürdige Art geholfen, indem sie nicht nur den Fragebogen ausgefüllt haben, sondern ausführliches Beschreibungsmaterial beileigten und bei Rückfragen bereitwillig weitere Auskünfte erteilten.

[**6]
Die entsprechenden Anmerkungen in (Hoepfner83a) lauten:

- L. Dreschler, H.H. Nagel, Volumetric model and 3D-trajectory of a moving car derived from monocular TV-frame sequences of a street scene. In: Proc. 7th IJCAI, Vancouver, 1981, pp 692-692.
 - B. Neumann, H. Marburger: Towards natural language description of real-world image sequences. In: J. Nehmer, (ed) GI - 12. Jahrestagung, Heidelberg 1982. p. 392-402

[**7]

DIN 2341. Die Deutsche Industrie(vor)norm 2341 wird abgekürzt MATER genannt und bedeutet "Magnetband Austauschformat für terminologische/lexikographische Daten". MATER ist im Normentaschenbuch 154 beschrieben (Vergl. Anmerkung (**11), p.84ff.

Der Geltungsbereich dieser Norm wird wie folgt umschrieben:

"Die Norm beschreibt die Anforderungen, die beim Austausch von ein- oder mehrsprachigen terminologischen und/oder lexikographischen Daten auf Magnetbändern von Benutzern von Datenverarbeitungsanlagen beachtet werden sollen. Der Austausch auf anderen Datenträgern wird in dieser Norm nicht behandelt. Unter terminologischen/lexikographischen Daten sollen in dieser Norm insbesondere verstanden werden:

- a) Wörterbücher und Lexika.
 - b) Dokumentationssprachen.
 Dazu werden gerechnet:
 - Klassifikationssysteme (z.B. Dezimalklassifikation DK)
 - Thesauri [s. DIN 1463]
 - Schlag- und Stichwortverzeichnisse.
 - c) Sonstige Wortsammlungen und Konkordanzen.
 - c) Sonstige Wortsammlungen und Konkordanzen.
 Dazu werden u.a. gerechnet:
 - terminologische Wortsammlungen
 - lexikographische Zusammenstellungen. "
- (p.85)

Der Zweck der Norm ist folgender:

"Diese Norm soll ihren Anwendern bei Verwendung kompatibler Kategorien den Austausch von terminologischen/lexikographischen Daten mit Hilfe der Datenverarbeitung ermöglichen. Jedem Ersteller von Austauschbändern wird die Gelegenheit gegeben, unabhängig vom jeweiligen Empfänger die Daten seiner Datenbank bzw. Teile davon in einheitlicher Form auf Magnetband zu speichern. Darüber hinaus gestattet es die Norm, jedem Empfänger von Austauschbändern verschiedener Ersteller, mit geringem Programmieraufwand die Daten zu übernehmen bzw. Teile davon auszuwählen.

Es ist nicht Zweck dieser Norm festzulegen, wie und in welcher Form ein Anwender seine Daten intern speichert. Das 'Magnetband-Austauschformat für terminologische/lexikographische Daten (MATER)' wurde speziell für den Austausch terminologischer und lexikographischer Daten entwickelt, da die für das Bibliothekswesen und für die Dokumentation u.a. vorliegenden Austauschformate 'Maschinelles Austauschformat für Bibliotheken (MAB1)' und 'Magnetband-Austauschformat für Dokumentationszwecke (MADOK)' hier nicht anwendbar sind. Soweit wie möglich wurden jedoch die physikalische Struktur und der logische Aufbau von MATER an die entsprechenden Teile von MAB1 bzw. MADOK angeglichen. Ebenso wurde versucht eine möglichst einheitliche Terminologie zu verwenden."

(p.85)

Aus DIN 2341 ging der internationale Normentwurf ISO/DP 6156 hervor.

Im Zusammenhang mit Datenaustausch auf Magnetbändern sind folgende Normen mitgültig:

"MATER regelt nur die Spezifikationen, die speziell für den Austausch terminologischer/lexikographischer Daten benötigt werden. Alle andern Regelungen für den Aufbau von Austauschbändern sind in folgenden DIN-Normen enthalten:

DIN 1506 Format für den Austausch von bibliographischen Daten
DIN 31624 Erweiterter Zeichenvorrat für bibliographische Daten bei Verwendung lateinischer Schriftzeichen
DIN 31626 Erweiterter Steuerzeichenvorrat für bibliographische Daten
DIN 66003 Informationsverarbeitung: 7-Bit-Code
DIN 66029 Kennsätze und Dateianordnung auf Magnetbändern für den Datenaustausch
DIN 66203 Informationsverarbeitung: 7-Bit-Code, Regeln zur Erweiterung"
(p.85)

[**8]
KBMS bedeutet Knowledge Based Managment System. Darauf bezieht sich auch der offizielle Titel des Projekts KIT-BACK "Advanced Data and Knowledge Management System". Das Projekt wird in Kooperation mit den Firmen Nixdorf, Olivetti und Bull sowie den Universitäten Turin, Bologna und Dortmund durchgeführt.

[**9]
"KL-ONE is a knowledge representation system developed at Bolt Beranek and Newman over the past few years (see [Brachman 77, Brachman 79, Schmolze 82, Sidner81]) that grew out of semantic network formalisms".
(Schmolze83, p.330)

Die von Schmolze erwähnten Publikationen sind:

- J.R. Brachman, A Structural Paradigm for Representing Knowledge, PH.D.thesis, Harvard University, May 1977. (Entspricht BBN Report No.3605, Bolt Beranek and Newman Inc, 1978)
- J.R. Brachman, R.J. Bobrow, P.R. Cohen, J.W. Klovstad, B.L. Webber, W.A. Woods, Research in Natural Language Understanding - Annual Report: 1.Sept. 78 - 31. Aug 79, Bolt Beranek and Newman Inc, BBN Report No. 4374, August 1979.
- J.G. Schmolze, Proceedings of the 1981 KL-ONE Workshop. Bolt Beranek and Newman Inc. BBN Report No. 4842, June 1982
- C.L. Sidner, M. Bates, R.J. Bobrow, R.J. Brachman, P.R. Cohen, D. Israel, J. Schmolze, B.L. Webber and W.A. Woods, Research in Knowledge Representation vor Natural Language Understanding, Annual Report: 1 September 1980 - 3. August 1981, Bolt Beranek and Newman Inc., Cambridge (Mass.), BBN Report No. 4785, 1981.

Vergl. auch:

R.J. Brachman, J.G. Schmolze, An Overview of the KL-ONE Knowledge Representation System. (bibliographiert als (Brachman85)). In einer Anmerkung nennen Brachman und Schmolze Namen von Forschern, die an

der Entwicklung von KL-ONE und Varianten des Formalismus beteiligt gewesen sind:

"At the very last, we would like to extend our thanks to these people, who have been intimately involved with KL-ONE, KL-TWO, and Krypton: Danny Bobrow, Rusty Bobrow, Phil Cohen, Richard Fikes, Mike Freeman, Jeff Gibbons, Victoria Gilbert, Brad Goodman, Norton Greenfield, Austin Henderson, David Israel, Henry Leitner, Hector Levesque, Bob Lingard, Tom Lipkis, Bill Mark, Peter Patel-Schneider, Candy Sidner, Mark Stefik, Marc Vilain, David Wilczynski, Bill Woods Martin Yonke, and Frank Zdybel"
(Brachman85, p.171)

[**10]

SRL bedeutet Semantic Representation Language. Diese Sprache wurde in Berlin bereits im BEAST-System verwendet (BEAST = Berlin Automatic Semantic-Oriented Translation). SRL dient dazu, natürlichsprachliche Ausdrücke semantisch formal zu repräsentieren:

"We want to give the following definition (*):

meanings of natural language expressions are well-formed expressions of a semantic representation language. Hence the assignment of meaning is a mapping from the natural language into the semantic representaton language:

NL -----> SRL

NL = natural language,
SRL = seamantic representation language."
(Habel80, p.61)

(*) Zu dieser Definition merken Habel et al. an:

"This definition does not (!) give an exact explanation of the term 'meaning'"
und:
"The 'equalization' of 'meaning' and 'expression of SRL', however, is sufficient for our further considerations and research."
(Habel80, p.61)

über neuere Entwicklungen in der Sprache SRL berichten:

- W. Emde, K.v.Luck und A. Schmiedel, Eine neue Implementation von SRL. In: Informatik Fachberichte 103, Heidelberg, 1985, p.219.

In (Rollinger84) wird SRL ebenfalls diskutiert.

[**11]

H.-W. Hein erwähnt an dieser Stelle in zwei Anmerkungen:

- H. Niemann, H.-W. Hein, A Program of Parallel Processes for Understanding Continuous Speech. Computing, Suppl. 3, p. 141-148, 1981
und
- H. Niemann, H.-W. Hein, Expert Knowledge for Automatic Understanding of Continuous Speech. In: M. Kunt, F. de Coulon (eds), Signal Processing: Theories and Applications, Amsterdam, 1980, p.647-651.

[**12]

H.-W. Hein erwähnt an dieser Stelle in einer Anmerkung:

- G.D. Forney, The Viterbi Algorithm. In: Proc. IEEE, Vol 61, No. 3, 1973, p.268-278.

[**13]

Diese Angaben entstammen einem Brief von Heinz-Dirk Luckhardt vom 15. 3. 1985.

[**14]

DIN MATER vergl. Anmerkg. [**7].

ANMERKUNGEN ZU TEIL 3.

[**1]

Die Fragen zur Implementation wurden nicht aufgrund systematischer, informatischer Strukturierung des Objektbereichs, sondern teilweise in in Vorkenntnis konkreter Züge einzelner Lexikonsysteme ausgearbeitet.

Wenn man den in dieser Fragengruppe angesprochenen Objektbereich taxonomisch gliedern will, ergibt sich bezüglich der Fragen c) und d) die Möglichkeit die Zusammenhänge und Ebenen übersichtlich zu machen, indem der Komplex "Datenbank, DBMS, Zugriffsalgorithmen" vereinfacht wie folgt dargestellt wird:

Datenbanksysteme können eingeteilt werden in:

- a) File-Systeme und Zugriffssysteme,
- b) Relationale Datenbanksysteme,
- c) Hierarchische Datenbanksysteme und
- d) Netzwerkartige Datenbanksysteme.

Diese erste Grobeinteilung, insbesondere die Unterscheidung zwischen relationalen, hierarchischen und netzwerkartigen Systemen ist üblich und findet sich z.B. in: G. Schlageter, W. Stucki, Datenbanksysteme: Konzepte und Modelle, Teubner, Stuttgart, 1981. Die Begriffe "relational", "hierarchisch" und "netzwerkartig" sind eindeutig und bedürfen keiner weiteren Erörterung.

Die unter a) genannte Klasse "File-Systeme und Zugriffssysteme" hingegen kann weiter differenziert werden. Eine solche Differenzierung trägt der Tatsache Rechnung, dass bestimmte Access-Methods (wie VSAM und ISAM) als Bündel von Zugriffsalgorithmen und Fileverwaltungsmethoden quasi-normiert sind, während andere Access-Methods eher auf spezifische und individuelle Bedürfnisse ausgerichtet sind. Innerhalb dieser zweiten Gruppe könnte wiederum differenziert werden, indem als Differenzierungskriterium das Merkmal "Zugriffsalgorithmus" verwendet wird. Im Sinne einer unvollständigen Illustration kann diese Einteilung bezüglich "Zugriffsalgorithmus" ansatzweise wie folgt gemacht werden:

- a) Nur sequentieller Zugriff,
- b) Indextechniken,
- c) B-Baum,
- e) Hashing-Verfahren.

Zusammenfassend dargestellt ergibt sich die folgende Struktur:

```

DATABANKSYSTEME
  relationale
  hierarchische
  netzwerkartige
FILE-SYSTEME und ZUGRIFFSSYSTEME
  VSAM, ISAM
  ANDERE und EIGENE ACCESS-METHODS
    nur sequentieller Zugriff
    Indextechniken
    B-Baum
    hashing

```

Diese Übersicht bildet lediglich eine Ergänzung zur Gruppe 3 des Fragebogens und hat rein illustrativen Charakter.

[**2]

"Die Kongruenzanalyse wird mit dem Prädikat

'combine(*Synt1,*Synt2,*Synt1_,*Synt2_).'

realisiert. Es verknüpft die beiden Listen von syntaktischen Merkmalsreihen *Synt1 und *Synt2 so, dass (in den Ergebnisvariablen *Synt1_ und *Synt2_) nur kongruente Reihen übrig bleiben." (Eimermacher83b, p.47)

Kongruenzen werden je zwischen den Wortartenpaaren

- VG/NG
- Praep/NG
- NG/NG
- NG/RelPron

abgeprüft. (Die verwendeten Abkürzungen für die Wortarten sind in Figur 20 erklärt). Die Kongruenzen betreffen:

- Kas
- Num
- Gen
- Pers
- Grad
- Stellung
- Temp
- Mod
- GenV
- FLEX

(Diese Abkürzungen sprechen für sich selbst).

[**3]

An dieser Stelle macht M. Eimermacher folgende Anmerkung:

"In der hier vorgestellten einfachen Version erkennt man Mehrdeutigkeiten nur dadurch, dass man durch ein 'fail' am Ende des

Parseraufrufs diesen zwingt, durch Backtrack weitere Repräsentationen für die Eingabe zu erstellen. Ausnahmen bestehen lediglich für verwandte Lesarten, d.h. solchen, die sich nicht in der Wortart unterscheiden."
(Eimermacher83c, p.11)

[**4]
"Bei Verben sind, um das gesamte Paradigma zu erfassen, maximal folgende Stämme bzw. Formen notwendig, z.B. 'geben', 'horchen':
(Benutzeranleitung zu SADAW , p.17)

Infinitivstamm	geb	horch
Stamm der 2.Pers.Sg.Präs	gib	---
Partizip II-Form	gegeben	gehört
Inf. mit "ZU" im Wortlaut	(abzugeben)	(aufzuhören)
Präteritumstamm	gab	hört
Konjunktiv II-Stamm	gaeb	---

(Benutzeranleitung zu SADAW , p.17)

[**5]
IAB 9 bedeutet interner Arbeitsbericht Nr. 9 und ist im Literaturverzeichnis als (Eimermacher83b) bibliographiert.

[**6]
Gemeint ist der Lehrstuhl für Informatik V (Mustererkennung) des Instituts für Mathematische Maschinen und Datenverarbeitung (IMMD) der Universität Erlangen-Nürnberg.

[**7]
Bibliographiert als (Endres86).

[**8]
Für neuere Ansätze im Zusammenhang mit MÜ siehe: H.J. Weber, Ein dreistufiges Lexikonsystem zur maschinellen Analyse von semantischen Textstrukturen. In: Textstrukturen und maschinelle Übersetzung, Arbeitsberichte des Projekts A3 des SFB100, TEMA Nr. 6.2, Saarbrücken 1984. In COAT (=Kohärenz-Analyse von Texten) werden die Lexika COMPLEX (=Lexikon für komplexe Prädikate), ALEX (=Argument Lexikon) und TRAP (=Transfer von Propositionen) verwendet. Das im SFB100 entwickelte Übersetzungssystem SUSY, welches u.a. auf SADAW zugreift, wurde weiterentwickelt. Vergl. dazu: H.-D. Maas, SUSY II Handbuch. Neue Folge, Heft 14, Saarbrücken, Univ. des Saarlandes, SFB100, 1985.

[**9]
Die in dieser und der nächsten Zeile verwendeten Abkürzungen für Wörterbücher oder Projekte

- LIMAS und
- PLIDIS

sind in der Tabelle zur Datenbasis des kumulierten Lexikons in Figur 14 erklärt, wobei das Lexikon des Projekts PLIDIS als MOLEX (=morphologisches Lexikon) aufgeführt ist. Die vier anderen Lexika figurieren nicht in jener Tabelle.

- HAM/RPM (= Hamburger Redepartner Modell) ist das Vorläuferprojekt von HAM-ANS. Das dort verwendete Lexikon war im Gegensatz zu dem in HAM-ANS verwendeten Wortlexikon, welches ein Stammformenlexikon ist, ein Vollformenlexikon. (Vergl. Kapitel 2.2.1.1: HAM). Das Projekt HAM/RPM ist in (Bolc80), p.119ff beschrieben.

- BACON (=Berlin Automatic CONstruction of Semantic Networks) ist ein Frage-Antwort-System, welches 1975 bis 1981 unter der Leitung von H.J. Schneider im Rahmen des Projekts "Automatische Erstellung semantischer Netze" an der Technischen Universität Berlin entwickelt wurde. (Vergl. Kapitel 2.2.1.2: KIT). In den gleichen Projektrahmen gehört das System BEAST (= Berlin Automatic Semantic-Oriented Translation), welches in (Bolc80), p.57ff beschrieben ist.

Die beiden Lexika

- DEUSEM und
- JUDO

gehören zum SFB100. DEUSEM heisst das vormalis SESAM genannte Deutsche Semantische Wörterbuch. Als SESAM ist es in (Eggers80), p.150ff beschrieben, wo es der semantischen Disambigulierung dient. Es enthält die folgenden drei Klassen von Einträgen:

- a) Valenzrahmen-Einträge:
Stichwort, Bedeutungsnummer, Valenzen (max. 4 Kasus) und die zu den Valenzen gehörenden Merkmalmengen.
- b) Merkmal-Einträge:
Stichwort, Bedeutungsnummer und Merkmale.
- c) Regel-Einträge:
Stichwort, Bedeutungsnummer und Regeln.

(Eggers80, p.150)

Eggers et al. erklären diese Eintragsklassen wie folgt:

"Die Valenzrahmen-Einträge bestimmen die Kasus und semantischen Subkategorien der notwendig abhängenden Komplemente; die Stichwörter sind Verben.

Die Merkmaleinträge liefern die semantischen Subkategorien für Substantive und Präpositionen.

Die Regel-Einträge enthalten Regeln, die als Elemente einer nicht-prozeduralen Sprache aufgefasst werden können. Die Regeln formulieren Kontextbedingungen in Form von Abfragen und Anweisungen. Folgende Abfragen sind derzeit möglich:

- Subklasse einer Nominalgruppe bzw. ihres Kerns (z.B. 'abstrakt', 'konkret', 'lokal')
- der lemmatisierte Wortlaut einer Nominalgruppe (z.B. GRÜN TISCH)
- der lemmatisierte Wortlaut des Kerns einer Nominalgruppe (z.B. TISCH)
- die Anzahl Wortformen einer Nominalgruppe
- der Numerus einer Nominalgruppe
- die Wortklasse des Kerns einer Nominalgruppe (z.B.

'SUB', 'DEM')

- das Vorkommen einer Wortklasse innerhalb einer Nominalgruppe (z.B. 'NUM')
- der lemmatisierte Wortlaut eines Adjektivs innerhalb der Nominalgruppe
- das Vorliegen von Reflexivität beim Verb."

(Eggers80, p.150)

Das Wörterbuch DEUSEM wird sowohl im Sprachübersetzungssystem SUSY als auch in abgewandelter Form im Dokumentationssystem JUDO benutzt.

Auf Fragen der Integration der Wörterbücher SADAW und DEUSEM im System Judo geht H. Werner ein: "Eine Sprachdatenbank: Zur Gestaltung der Wörterbucharbeit" (Werner80). Vergl. dazu auch: (Zimmermann78).

[**10]

Der Transformationsprozess des Semantik-Moduls

Syntax-->Pragmatik

ist in mehrere Teilprozesse aufgegliedert, bei denen Lexikoninformation verwertet wird. Diese Teilprozesse sind: (Der nun folgende Überblick basiert auf Brietzmann84a, p77-144)

1. Lokale Interpretation: Amalgamierung und Kasusrahmenspezialisierung.

Diese Funktion heisst KONSTITUENTEN_INTERPRETATION und wird von den Hilfsfunktionen

-AMALGAM,

-VERBFRAME_GENERIERUNG und

-SEMANTIK_HYPOTHESEN_GENERIERUNG

unterstützt.

2. Kontextuelle Interpretation: Kasusrahmenanalyse. (Inklusive Kasusanalyse freier Angaben).

Die hier benötigten Hauptfunktionen sind:

-FRAME_SATZ_MATCH,

-FRAME_KONSTIT_MATCH und

-INTERPRETATION_VERVOLLSTÄNDIGEN.

3. Analysesteuerung durch Kontexterwartungen weiterer Erkennungsergebnisse aus Zwischenergebnissen.

Diese Funktion heisst

-TOP_DOWN_HYPOTHESEN

und dient der Generierung von Hypothesen, die als Erwartungen aus Teilergebnissen abgeleitet werden können.

[**11]

Zum Verhältnis zwischen syntaktischer und semantischer Valenz vergl. (Helbig82), p.10 und p.62 ff.

VERWENDETE LITERATUR

(Appelrath85)

Hans-Jürgen Appelrath, Von Datenbanken zu Expertensystemen. Informatik Fachberichte 102, Heidelberg, 1985.

(Ballweg79)

Angelika Ballweg-Schramm, Helmut Schumacher, Verbalenzwörterbuch auf semantischer Basis. In: Helmut Henne, Praxis der Lexikographie. Berichte aus der Werkstatt, Tübingen, 1979, p. 94.

(Berger85)

Dieter Berger, Günther Drosdowski, et al. (hg) Duden. Richtiges und gutes Deutsch, Wörterbuch der sprachlichen Zweifelsfälle, Duden Bd. 9, Mannheim 1985.

(Bergmann82)

Henning Bergmann, Lemmatisierung in HAM-Ans. In: Memo Ans-10, Forschungsstelle für Informationswissenschaft und Künstliche Intelligenz, Universität Hamburg, 1982.

(Billmeier78)

Reinhard Billmeier, Konzeption einer automatischen Inhalts- und Strukturanalyse für ein Informations-Retrieval-System. in: SDV, 2. Jahrgang, Heft 2, p103.

(Bolc80)

L. Bolc (hg), Natural Language Based Computer Systems, München, 1980

(Bondzio82)

Wilhelm Bondzio, Valenz in der Lexikographie. In: E. Agricola, J. Schildt, D. Viehweger, Wortschatzforschung heute. Aktuelle Probleme in der Lexikologie. Linguistische Studien. Leipzig, 1982, p.127.

(Boot79)

Martinus Boot, Homographie. Ein Beitrag zur automatischen Wortklassenzuweisung in der Computerlinguistik. Bilthoven, 1979.

(Brachman77)

Ronald J. Brachman, A Structural Paradigm for Representing Knowledge. Ph. D. Thesis, Harvard, 1977.

(Brachmann79)

Ronald J. Brachman, On the epistemological Status of Semantic Networks. In: N. Findler (hg.) Associative Networks. Representation and Use of Knowledge by Computers. New York, 1979.

(Brachmann85)

Ronald J. Brachman, James G. Schmolze, An overview of the KL-One Knowledge Representation System. In: Cognitive Science 9, 1985, p.171-216.

(Bredemeyer77)

Jürgen Bredemeyer, Valenztheorie und Kasusgrammatik. In: (Petöfi77), p. 427

(Brietzmann84a)

Astrid Brietzmann, Semantische und pragmatische Analyse im Erlanger Spracherkennungsprojekt. Diss. In: Arbeitsberichte des Instituts für Mathematische Maschinen und Datenverarbeitung, Band 17, Nr. 5, 1984.

(Brietzmann84b)

Astrid Brietzmann, Eine Netzwerkgrammatik des Deutschen für die automatische Spracherkennung. In: Sprache und Datenverarbeitung, 8. Jahrgang, Heft 1/2, 1984, p.54.

(Brun86)

Georg Brun, Thomas Schnurrenberger, Analyse des Wörterbuchs zur Valenz und Distribution deutscher Verben von G. Helbig und W. Schenkel im Hinblick auf eine Implementation in maschinenlesbarer Form. Seminararbeit, Zürich, 1986.

(Brustkern83)

Jan Brustkern, W. Schulze, Aufbau einer Wortdatenbank für das Deutsche in: W. Lenders/H. Zimmermann (hg.), Sprache und Datenverarbeitung, Sonderdruck SDV 2/1980, Tübingen 1983.

(Busemann84)

Stephan Busemann, Surface Transformations During the Generation of Written German Sentences, HAM-ANS Report Nr.27, Hamburg, 1984.

(Date86)

C.J. Date, An Introduction to Database Systems. Vol I, Reading, Mass., 1986(4).

(Date83)

C.J. Date, An Introduction to Database Systems. Vol II, Reading, Mass., 1983

(Denisov82)

P.N. Denisov, Über die universelle Struktur des Wörterbuchartikels. In: Werner Wolski (hg.), Aspekte der sowjetrussischen Lexikographie. In: Reihe Germanistische Linguistik 43 (hg. Helmut Henne, Horst Sitta, Herbert Ernst Wiegand), Tübingen, 1982, p. 89.

(Dietrich73)

Rainer Dietrich, Automatische Textwörterbücher, Studien zur maschinellen Lemmatisierung verbaler Wortformen des Deutschen. Tübingen, 1973.

(Dietrich74)

Rainer Dietrich, Wolfgang Klein, Computerlinguistik. Eine Einführung. Stuttgart/Berlin/Köln/Mainz 1974

(Domenig86)

Marc Domenig, Entwurf eines dedizierten Datenbanksystems für Lexika, Zürich, 1986

(Dornseiff33)

F. Dornseiff, Der Deutsche Wortschatz nach Sachgruppen. Berlin 1933.

(Drewek75)

Raimund Drewek, LDVLIB - Dokumentation der Programme zur computerunterstützten Textanalyse, Zürich 1975 und später.

(Drewek80/83)

Raimund Drewek, LEM, Interaktive Wörterbuchunterstützte
Lemmatisierung. Stuttgart/Zürich 1980/83.

(Drewek82)

R. Drewek, O. Zäch, LDVLIB - Glossar zur LDV-Terminologie.
Stuttgart/Zürich, 1982.

(Drewek84)

Raimund Drewek, NET/1 - Netzwerkerzeugte Typographie für
Wörterbücher, Register und Kataloge. Zürich/Stuttgart, 1984.

(Drewek85)

Raimund Drewek, Vorbemerkungen zu einer computergestützten Lösung
für die Lexikographie im Verlag als Beitrag zur praktischen SDV.
In: Informatik-Fachberichte 114, Heidelberg, 1985, p.34 .

(Drewek86)

Raimund Drewek, EDP and Lexicography. Referat beim IV. Forum
Information Science and Practice 'Standardization in computerized
Lexicography', Saarbrücken, Oktober 1986. (Org. Deutsche
Forschungsgemeinschaft).

(Drosdowski77)

Günther Drosdowski, Helmut Henne, Herbert E. Wiegand, Nachdenken
über Wörterbücher. Mannheim, 1977.

(Drosdowski79)

Günther Drosdowski, Die Metapher im Wörterbuch. In: (Ballweg79) ,
p. 83.

(Eggers69)

Hans Eggers, R. Dietrich, R. Rath, A. Rothkegel, H.J. Weber, H.
Zimmermann, Elektronische Syntaxanalyse der deutschen
Gegenwartssprache. Tübingen, 1969.

(Eggers80)

Hans Eggers, Heinz-Dirk Luckhardt, Heinz-Dieter Maas, Monika
Weissgerber (zusammengestellt), SALEM. Ein Verfahren zur
automatischen Lemmatisierung deutscher Texte, Tübingen, 1980.

(Ehrlich83)

Ute Ehrlich, Semantische Klassifikation und Valenzbeschreibung
deutscher Adjektive und Substantive für das Lexikon der
Automatischen Sprachverarbeitung. Studienarbeit Informatik.
Betreuerin: Astrid Brietzmann. Erlangen/Nürnberg 1983.

(Ehrlich86)

Ute Ehrlich, Ein Lexikon für das natürlichsprachliche Dialogsystem
EVAR. In: Arbeitsberichte des Instituts für Mathematische Maschinen
u. Datenverarbeitung, Bd. 19, Nr. 3, Universität Erlangen/Nürnberg,
1986.

(Eimermacher83a)

Michael Eimermacher, Ein Word Expert Parser in PROLOG - Integration
von Syntax und Semantik. In: KIT Interner Arbeitsbericht 8 (IAB8),
TU, Berlin, 1983.

(Eimermacher83b)

Michael Eimermacher, Ein Word Expert Parser in PROLOG - Die

lexikalische Komponente. In: KIT Interner Arbeitsbericht 9 (IAB9), TU, Berlin, 1983.

(Eimermacher83c)

Parsing in PROLOG - Eine semantische Komponente. In: KIT-Report 15. TU Berlin, 1983.

(Eimermacher85)

Brief vom 20.3. 1984 als Antwort auf Fragebogen.

(Emde82)

W. Emde, C. Habel, C.-R. Rollinger, Automatische Akquisition von inferentiellem Wissen. In: Informatik Fachberichte 58, Heidelberg, 1982, p.72.

(Emde85)

W. Emde, K. Luck, A. Schmiedel, Eine neue Implementation von SRL. In: Informatik Fachberichte 103, Heidelberg, 1985, p.219.

(Endres86)

Hermann Endres, Ein Datenverwaltungssystem für ein Lexikon zur automatischen Spracherkennung und Sprachgenerierung. Diplomarbeit Informatik. Betreuerin: Ute Ehrlich. Erlangen/Nürnberg, 1986.

(Eroms85)

Hans Werner Eroms, Eine reine Abhängigkeitsgrammatik für das Deutsche. In: Deutsche Sprache Heft 4/1985, p. 306.

(Fillmore71)

Charles J. Fillmore, Some Problems of Case Grammar, Ohio, 1971.

(Guenthner/Lehmann86)

Franz Guenthner, Hubert Lehmann, Verarbeitung natürlicher Sprache - ein Überblick. In: Informatik Spektrum, Bd. 9, Heft 3, Heidelberg, 1986, p. 162.

(Guckler77)

Gudrun Guckler, Evelyn Müller, Gerhard Wahrig, Wörterbuch als Datenbank. Ein komplexes linguistisches, logisches und technisches Problem. In: SDV 1. Jahrgang 1977 Heft 2, p.105.

(Habel78)

C.U. Habel, A. Schmidt, H. Schweppe, BEAST - Ein natürlichsprachliches Datenbankinterface. in: SDV 2. Jahrgang, Heft 2, 1978, p.121.

(Habel80)

C.U. Habel, Claus-Rainer Rollinger, Arno Schmidt, Hans-Jochen Schneider, A logic-oriented approach to automatic text understanding. In: Leonard Bolc (ed), Natural Language Based Computer Systems. München/Wien 1980, p.57

(Habel81)

C.U. Habel, C.-R. Rollinger, Termsortierung in SRL. In: Bericht 80/1 Projekt Semantische Netze. TU Berlin, 1981

(Habel82a)

C.U. Habel, C.-R. Rollinger, EZH-Strukturen als Textrepräsentation - Ein Beispiel. In: Istvan Bátori, Jürgen Krause, Hans Dieter Lutz (hg), Linguistische Datenverarbeitung, Tübingen, 1982, p. 137).

(Habel82b)

C.U. Habel, Probleme und Konzepte der maschinellen Sprachverarbeitung aus Sicht der Künstlichen Intelligenz und Cognitive Science. In: Istvan Bátori, Jürgen Krause, Hans Dieter Lutz (hg), Linguistische Datenverarbeitung, Tübingen, 1982, p. 39.

(Habel85a)

C.U. Habel, Das Lexikon in der Forschung der Künstlichen Intelligenz. In: Ch. Schwarze, D. Wunderlich (hg), Handbuch der Lexikologie, Königstein/Ts., p.444.

(Habel85b)

C.U. Habel, C.-R. Rollinger, Lernen und Wissensakquisition. in: Informatik Fachberichte 93. Heidelberg, 1985, p. 249.

(Hahn85a)

Udo Hahn, Informationslinguistik I, Einführung in das linguistische Information Retrieval. Kursskript Lehrveranstaltung WS1982/83, Konstanz, 1985.

(Hahn85b)

Udo Hahn, Informationslinguistik I, Linguistische und statistische Verfahren im experimentellen Information Retrieval. Kursskript Lehrveranstaltung. SS1983/84, Konstanz, 1985.

(Hahn85c)

Udo Hahn, Informationslinguistik III, Natürlichsprachliche Systeme. In: Udo Hahn, Intelligente Informationssysteme. Konstanz, 1985, p.236.

(v. Hahn75)

W.v. Hahn, D. Henskes, W. Hoepfner, W. Wahlster, HAM-RPM: Ein Redepartnermodell als Simulationsprogramm. In: K. Braunnüller, W. Kürschner (hg.), Grammatik. Akten des 10. Linguistischen Kolloquiums Tübingen 1975, Bd. 2, Tübingen, 1976, p.337-357.

(v. Hahn80)

Walther v. Hahn, Wolfgang Hoepfner, Antony Jameson, Wolfgang Wahlster, The Anatomy of the Natural Language Dialogue System HAM-RPM. In: Leonard Bolc (ed) Natural Language Based Computer Systems. München/Wien, 1980 (Bolc80), p119.

(Hausmann85)

Franz Josef Hausmann, Lexikographie. In: Ch. Schwarze, D. Wunderlich (hg), Handbuch der Lexikologie, Königstein/Ts., p.367.

(Heger71)

Klaus Heger, Monem, Wort und Satz. In: Konzepte der Sprach- und Literaturwissenschaft. Tübingen, 1971.

(Heger84)

Klaus Heger, Klaus Muderbach, Aktantenmodelle. Aufgabenstellung und Aufbauregelung. In: Abhandlungen der Heidelberger Akademie der Wissenschaften. Philosophisch-historische Klasse. Jahrgang 1984, 4. Abhandlung. Heidelberg, 1984.

(Heger85)

Klaus Heger, Flexionsformen, Vokabeln und Wortarten. In: Abhandlungen der Heidelberger Akademie der Wissenschaften. Philosophisch-historische Klasse. Jahrgang 1985, 1. Abhandlung.

Heidelberg, 1985.

(Hein81)

H.-W. Hein, A System Understanding Continuous German Speech. In: Informatik-Fachberichte 47, Heidelberg, 1981, p.277 .

(Hein82)

H.-W. Hein, Das Erlanger Spracherkennungssystem, Dissertation Universität Erlangen-Nürnberg, Arbeitsberichte des IMMD Band 15, 1982.

(Hein83)

H.-W. Hein, Das Arbeitstreifen 'Spezifikation einer lexikalischen Wissensbasis (Deutsch) für natürlichsprachliche KI-Systeme' In: Rundbrief der Fachgruppe Künstliche Intelligenz in der Gesellschaft für Informatik, Nr. 30, Juni 1983, p28.

(Heinze85)

G. Heinze, J. Brustkern, das Lexikonsystem ALEXSYS. In: IKP Arbeitsberichte Nr.6, Abteilung LDV, Bonn, 1985, p.56.

(Helbig69) (=Helbig75)

G. Helbig, W. Schenkel, Wörterbuch zu Valenz und Distribution deutscher Verben. Leipzig, 1969.

(Helbig82)

Gerhard Helbig, Valenz, Satzglieder, semantische Kasus, Satzmodelle. Leipzig, 1982.

(Helbig83)

Gerhard Helbig, Studien zur deutschen Syntax. Bd. 1. Linguistische Studien. Leipzig, 1983.

(Helbig84)

Gerhard Helbig, Studien zur deutschen Syntax. Bd. 2. Linguistische Studien. Leipzig, 1984.

(Hendrix78)

G.G. Hendrix, Semantic aspects of translation. In: D.E. Walker (ed), Understanding Spoken Language. New York, 1979, p.193.

(Henne79)

Praxis der Lexikographie, Tübingen, 1979.

(Herberg82)

Dieter Herberg, Neuere Erkenntnisse zu den Strukturprinzipien von Wortbedeutungen und ihre Widerspiegelung in Wörterbüchern. In: E. Agricola, J. Schildt, D. Viehweger, Wortschatzforschung heute. Aktuelle Probleme in der Lexikologie. Linguistische Studien. Leipzig, 1982, p.149.

(Heringer84)

Hans Jürgen Heringer, Kasus und Valenz. Eine Mésalliance? In: Zeitschrift für germanistische Linguistik, 12.2. Berlin/ NewYork, 1984, p.200.

(Hess83)

Klaus Hess, Jan Brustkern, Winfried Lenders, Maschinenlesbare deutsche Wörterbücher. Dokumentation, Vergleich, Integration. Tübingen, 1983.

(Heuer83)

Walter Heuer, Richtiges Deutsch. Neu bearbeitet von Max Flückiger und Peter Gallmann. Zürich, 1983.

(Hoepfner82)

W. Hoepfner, ATN-Steuerung durch Kasusrahmen. In: Informatik Fachberichte 58, Heidelberg, 1982, p.215.

(Hoepfner83a)

W. Hoepfner, Th. Christaller, H. Marburger, K. Morik, B. Nebel, M. O'Leary, W. Wahlster, Beyond Domain-Independence: Experience with the Development of a German Language Access System to Highly Diverse Background Systems. In: A. Bundy (ed), ijcai-83, Proceedings of the Eighth International Joint Conference on Artificial Intelligence (8-12 August 1983, Karlsruhe), Los Altos(Cal), 1983, vol. 1, p.588.

(Hoepfner83b)

Wolfgang Hoepfner, Lexikalische Wissensquellen im Dialogsystem HAM-ANS. Arbeitspapier Nr. 15 der Forschungsstelle für Künstliche Intelligenz, Uni Hamburg, 1983.

(Hohnhold83)

Hohnhold, TEAM-Eintragsschema - Stand 08/1983, Siemens Sprachendienst. München, 1983.

(Johnson83)

R. Johnson, Parsing with Transition Networks. In: M. King (ed), Parsing Natural Language. London etc, 1983.

(Imhasly82)

Bernhard Imhasly, Bernhard Marfurt, Paul Portmann, Konzepte der Linguistik. Wiesbaden, 1982.

(Kaiser84)

Stephan Kaiser, Der neue Dornseiff, Konzept einer Neubearbeitung von Franz Dornseiffs 'Der deutsche Wortschatz nach Sachgruppen'. In: ZGL(Zeitschrift für Germanistische Linguistik), 12.2. Berlin/NewYork, 1984.

(Kamp81)

Hans Kamp, A Theory of Truth and Semantic Representation. In: Groenendijk, Formal Methods in the Study of Natural Language. Amsterdam, 1981.

(Kilbury85)

James Kilbury, Chart Parsing and the Earley Algorithm. In: KIT-Report 24, TU Berlin, 1985

(Koskenniemi83)

K. Koskenniemi, Two-Level Morphology: A General Computational Model for Word-Form-Recognition and Production. Diss. Uni Helsinki, Dpt. of General Linguistics, Publication No. 11, 1983.

(Kroupa82)

E. Kroupa, Strategie und Dokumentrepräsentation bei CTX. Ein Verfahren zur computerunterstützten Texterschließung und Textwiedergewinnung. In: Istvan Bátori, Jürgen Krause, Hans Dieter Lutz (hg), Linguistische Datenverarbeitung, Tübingen, 1982, p. 156.

(Kühn)

Peter Kühn, deutsche Wörterbücher, Eine systematische Bibliographie. In: Reihe Germanistische Linguistik 15.

(Kühn79)

Peter Kühn, Der Grundwortschatz, Bestimmung und Systematisierung. In: Reihe germanistische Linguistik 17, 1979.

(Laisiepen80)

Klaus Laisiepen, Ernst Luterbeck, Karl-Heinrich Meyer-Uhlenried, Grundlagen der praktischen Information und Dokumentation. München 1980.

(Lamersdorf85)

Winfried Lamersdorf, Semantische Repräsentation komplexer Objektstrukturen. Modelle für nichtkonventionelle Datenbankanwendungen. Informatik fachberichte 100. Heidelberg, 1985.

(Laubsch85)

Techniken der Wissensdarstellung. In: Informatik Fachberichte 93. Heidelberg, 1985, p.48.

(Lehmann77)

Hubert Lehmann, Nikolaus Ott, Magdalene Zoeppritz, Benutzerspezifische Sprachen. In: SDV, 1. Jahrgang 1977, Heft 2, p. 127.

(Lehmann85)

Hubert Lehmann, Ein juristisches Expertensystem auf der Grundlage von Linguistik und Logik. In: Informatik-Fachberichte 114, Heidelberg, 1985, p.92 .

(Li84)

Deyi Li, A Prolog Database System. Letchworth, 1984.

(Lightning85)

Turbo Lightning, Owner's Handbook, Scotts Valley, 1985.

(Lutzeier85)

Peter Rolf Lutzeier, Die semantische Struktur des Lexikons. In: Christoph Schwarze, Dieter Wunderlich, Handbuch der Lexikologie. Königstein/Ts., 1985, p.103.

(Maas82)

H.-D. Maas, Repräsentation und Strategie bei der automatischen Analyse mit SUSY. In: Istvan Bátori, Jürgen Krause, Hans Dieter Lutz (hg), Linguistische Datenverarbeitung, Tübingen, 1982, p. 105.

(Machrone85)

Bill Machrone, Paul Somerson, Lightning Strikes. A Spark of Lightning. In: PC Magazine, December 10, 1985, p.112.

(Messerschmidt84)

Jan Messerschmidt, Linguistische Datenverarbeitung mit COMSKKEE. Stuttgart, 1984.

(Mresse84)

Moscheh Mresse, Information Retrieval - Eine Einführung, Stuttgart, 1984.

(Müller85)

Wolfgang Müller et al., Duden. Bedeutungswörterbuch. Wortbildung und Wortschatz. Mannheim 1985.

(Müller-Meernach81)

R. Müller-Meernach, Datenbasis für den Lexikonexperten eines Spracherkennungssystems (Studienarbeit im Fach Informatik). Betreuer: H.-W. Hein. Erlangen, 1981.

(Neubauer80a)

Fritz Neubauer, Janos S. Petöfi, Wortsemantik, Lexikonsysteme und Textinterpretation. In: Wolfgang Heydrich (hg) Lexikoneinträge. Grundelemente der semantischen Struktur von Texten V. In: Papiere zur Textlinguistik. Bd.31. Hamburg, 1980, p.149.

(Neubauer80b)

Fritz Neubauer, Die Struktur der Explikationen in deutschen einsprachigen Wörterbüchern. Eine vergleichende lexiko-semantische Analyse. Hamburg, 1980.

(Niemann82)

H. Niemann, The Erlangen System for Recognition and Understanding of Continuous German Speech. In: J. Nehmer (ed), GI - 12. Jahrestagung Informatik Fachberichte 57, Berlin, 1982

(Niemann85)

H. Niemann, The Speech Understanding and Dialog System EVAR. In: R. de Mori (ed), New Systems and Architectures for Automatic Speech Recognition and Synthesis. NATO-Asi Series 1985.

(Nida58)

A. Nida, Analysis of Meaning and Dictionary Making. 1958. In: Ladislav Zgusta, Probleme des Wörterbuchs, Darmstadt, 1985, p.21.

(Novak82)

H.-J. Novak, On the Selection of Verbs for Natural Language Description of traffic Scenes. In: Informatik Fachberichte 58, Heidelberg, 1982, p.22.

(Petöfi77)

Janos Petöfi, das lexikon in der grammatik die grammatik im lexikon (2 halbbände) Hamburg, 1977.

(Petöfi78)

Die Lexikonkomponente der TeSWest. In: Claudia Biasci, Johannes Fritsche (hg), Texttheorie, Textrepräsentation. Theoretische Grundlagen der kanonischen sinnsemantischen Repräsentation von Texten. Papiere zur Textlinguistik Bd.18. Hamburg, 1978, p.193.

(Petöfi80)

J. S. Petöfi, Explikationen in umgangssprachlichen und fachsprachlichen Wörterbüchern. In: Hans Jürgen Eikmeyer, Louise Jansen (hg), Objektargumente. Grundlagen der semantischen Struktur von Texten III. Papiere zu Textlinguistik. Bd.26, Hamburg, 1980, p.123.

(Piotrowski85)

R.G. Piotrowski, A.W. Popeskul, M.S. Chazinskaja, N.B. Rachubo, Automatische Wortschatzanalyse. In: Quantitative Linguistics, Vol.28, Bochum, 1985.

(Primio83)

Franco di Primio, Thomas Christaller, A Poor Man's Flavor System. (ed. M. Rosner). Working Paper No.4, Institut dalle Molle, Université de Genève, 1983.

(Quillian67)

M. Ross Quillian, Word Concepts: A Theory and Simulation of some Basic Semantic Capabilities. In: R.J. Brachman, H.J. Levesque (eds.), Readings in Knowledge Representation, Los Altos, 1985, p.98 ff.

(Rollinger84)

C.R. Rollinger (hg.), Probleme des (Text-) Verstehens. Ansätze der Künstlichen Intelligenz. Tübingen, 1984.

(Rollinger85)

C.R. Rollinger, Die Projektgruppe KIT an der TU Berlin. In: Rundbrief des Fachausschusses 1.2 Künstliche Intelligenz und Mustererkennung der Gesellschaft für Informatik (BRD), Nr.37, 1975, p.27 .

(Rostek82)

L. Rostek, Verstehen und Verdichten von Texten. Repräsentation und Strategie bei der Verarbeitung in einem Modell für ein Referenz-Retrieval-System, basierend auf einem Netz von Begriffsbezeichnungen. In: Istvan Bátori, Jürgen Krause, Hans Dieter Lutz (hg), Linguistische Datenverarbeitung, Tübingen, 1982, p. 127.

(Rostek85a)

Lothar Rostek, Dietrich Fischer, Konzepte für ein thesaurusbasiertes Information Retrieval am Arbeitsplatz. In: Sprachverarbeitung in Information und Dokumentation. Informatik Fachbericht 114. Heidelberg, 1985.

(Rostek85b)

Lothar Rostek, Dietrich Fischer, An Author's Workstation. Vision, Views and Activities. In: Protext II, Proceedings of the Second International Conference on Text-Processing Systems (October 1985, Dublin, Ireland). Separatdruck der Gesellschaft für Information und Dokumentation mbH (GID).

(Rohrer86)

Christian Rohrer, Maschinelle Übersetzung mit Unifikationsgrammatiken. In: István Bátori, Heinz J. Weber (hg), Neue Ansätze in Maschineller Sprachübersetzung: Wissensrepräsentation und Textbezug. Tübingen, 1986, p.75.

(Scerba82)

Versuch einer allgemeinen Theorie der Lexikographie. In: Werner Wolski (hg.), Aspekte der sowjetrussischen Lexikographie. In: Reihe Germanistische Linguistik 43 (hg. Helmut Henne, Horst Sitta, Herbert Ernst Wiegand), Tübingen, 1982, p. 17.

(Schefe82)

Peter Schefe, Some Fundamental Issues in Knowledge Representation. In: Informatik Fachberichte 58. Heidelberg, 1982, p.42.

(Schippan84)

Thea Schippan, Lexikologie der deutschen Gegenwartssprache. Leipzig 1984.

(Schmolze83)

James G. Schmolze, Thomas A. Lipkis, Classification in the KL-ONE System. In: A. Bundy (ed), ijcai-83, Proceedings of the Eighth International Joint Conference on Artificial Intelligence (8-12 August 1983, Karlsruhe), Los Altos(Cal), 1983, vol. 1, p.588.

(Schott78)

G. Schott, Automatische Kompositazerlegung mit einem Minimalwörterbuch zur Informationsgewinnung aus beliebigen Fachtexten. In: S. Koller, P.L. Reichertz, K. Überla, Klartextverarbeitung. Heidelberg, 1978, p.32.

(Schulz79)

J. Schulz, Eine Terminologiedatenbank für Übersetzer, TEAM. Sonderdruck aus PHILIPS terminology bulletin Vol.8 No.1/2, August 1979.

(Schulz81)

J. Schulz, Übersetzungshilfen im Dialog mit einer mehrsprachigen Terminologie-Datenbank. Abfragemöglichkeiten über Terminals. In: Babel, Separatum No 2/1981 Vol XVII. Revue Internationale de la Traduction.

(Schulze83)

Wolfgang Schulze, Gerlind Heinze, Die morphologische Komponente in der Wortdatenbank des Deutschen. In: SDV 82/1.2, p. 34 ff

(Schulze/Willée84)

Wolfgang Schulze, Gerd Willée, noch eine Variation über das Thema 'Lemma'. In: Sprache und Datenverarbeitung 1/83, Tübingen, 1984.

(Schumacher85)

Helmut Schumacher, Grammatik im semantisch orientierten Valenzwörterbuch. In: Henning Bergenholz, Joachim Mugdan, Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch. Tübingen, 1985.

(Schwind84)

Camilla B. Schwind, Natural Language Access to PROLOG Database Systems. In: Informatik Fachberichte 58. Heidelberg, 1982, p.237.

(Schwind85)

Camilla B. Schwind, Semantikkonzepte in der künstlichen Intelligenz. In: Informatik Fachberichte 93, Heidelberg, 1985, p.95.

(Sommerfeldt83a)

K.-E. Sommerfeldt, H. Schreiber, Wörterbuch zur Valenz und Distribution der Substantive, Leipzig, 1983.

(Sommerfeldt83b)

K.-E. Sommerfeldt, H. Schreiber, Wörterbuch zur Valenz und Distribution deutscher Adjektive, Leipzig, 1983.

(Steinbauer85)

Dieter Steinbauer, Hartmut Wedekind, Integritätsaspekte in Datenbanksystemen. In: Informatik Spektrum, Bd. 8, Heft 2, Berlin, April 1985, p.60.

(Stepanova81)

M.D. Stepanowa, G. Helbig, Wortarten und das Problem der Valenz in der deutschen Gegenwartssprache. Leipzig 1981.

(Teubert79)

Wolfgang Teubert, Valenz des Substantivs - Attributive Ergänzungen und Angaben, Düsseldorf, 1979

(Teubert84)

Wolfgang Teubert, Setting up a Lexicographic data-base for German. In: R.R.K. Hartmann, LEXeter Proceedings, Tübingen, 1984, p.425.

(Thiel82)

Manfred Thiel, Zur Verwendung der Kasusgrammatik im faktenorientierten 'Information Retrieval' in Textdaten. Tübingen, 1982.

(Trost84)

Harald Trost, Wissensrepräsentation in der AI am Beispiel semantischer Netze. In: Retti et al., Artificial intelligence. Stuttgart, 1984.

(Vetter82)

Max Vetter, Aufbau betrieblicher Informationssysteme. Stuttgart 1982.

(Vollnhals82)

Otto Vollnhals, Technical Dictionaries Retrieved from a Database. In: Meta, Journal des Traducteurs, Vol. 27, Nr.3, Montréal, 1982

(Vollnhals84)

Otto Vollnhals, Utilization of a Commercial Linguistic Data-base System for Electronic Storage and Automated Production of Dictionaries. In: R.R.K. Hartmann, LEXeter Proceedings, Tübingen, 1984, p.430.

(Wahlster82a)

W. Wahlster, Natürlichsprachliche Systeme. In: Informatik Fachberichte 59. Heidelberg, 1982, p.203.

(Wahlster82b)

W. Wahlster, Aufgaben, Standards und Perspektiven sprachorientierter KI-Forschung. Einige Überlegungen in informatischer Sicht. In: Istvan Bátor, Jürgen Krause, Hans Dieter Lutz (hg), Linguistische Datenverarbeitung, Tübingen, 1982, p. 13.

(Wahrig83)

Gerhard Wahrig, Gesammelte Schriften, Tübingen 1983.

(Wahrig85)

Gerhard Wahrig, Deutsches Wörterbuch, 1985.

(Weber84)

H.J. Weber, Semantische Textstruktur-Analyse mit COAT. In: A. Rothkegel, B. Sandig (hg), Text - Textsorten - Semantik. Hamburg, 1984, p.222.

(Weber85)

H.J. Weber, Coherence-oriented Lexican Structures and Text-based Semantic Representation of Lexemes. An Outline of COAT. In: G.A.J.

Hoppenbrowsers, P.A.M. Seuren, A.J.M.M. (N.N.O.P.) Weijters, Meaning and the Lexicon. Dordrecht 1985, p.317.

(Wegner85)

I. Wegner, Frame-Theorie in der Lexikographie. Untersuchung zur theoretischen Fundierung und computergestützten Anwendung kontextueller Rahmenstrukturen für die lexikographische Repräsentation von Substantiven. In: Lexicographica, Series Maior, Band 10, Tübingen 1985.

(Weinreich76)

Harald Weinreich, Die Wahrheit der Wörterbücher (1975/76). In: Ladislav Zgusta, Probleme des Wörterbuchs, Darmstadt, 1985, p.248.

(Werner80)

Helmut Werner, Eine Sprachdatenbank: Zur Gestaltung der Wörterbucharbeit. In: SDV 4. Jahrgang, Heft 1, 1980, p.88.

(Wersig78)

Gernot Wersig, Thesaurus Leitfaden. München 1978.

(Wiegand83)

Herbert E. Wiegand, Was ist eigentlich ein Lemma? ein Beitrag zur Theorie der lexikographischen Sprachbeschreibung. In: Germanistische Linguistik 1-4/82, p.401.

(Willée79)

Gerd Willée, LEMMA, ein Programmsystem zur automatischen Lemmatisierung deutscher Wortformen. In: SDV 1. Jahrgang, Heft 2, (1977), 1979, p.160.

(Willée84)

Gerd Willée, Das Programmsystem LEMMA2 - eine Weiterentwicklung von LEMMA. IKP Arbeitsbericht, Abt. LDV, Nr.2, Bonn, 1984.

(Willée85)

Gerd Willée, Das Bonner Lexikonsystem BONNLEX1. In: IKP Arbeitsberichte Nr. 6, Bonn 1985, p.1.

(Winograd83)

Language as a Cognitive Process, Vol 1, Syntax, Reading mass., 1983.

(Winograd84)

Terry Winograd, Software für Sprachverarbeitung. In: Spektrum der Wissenschaft, 11/1984, p.88.

(Wolski82)

Werner Wolski (Hg.), Aspekte der sowjetrussischen Lexikographie, Tübingen, 1982.

(Wüster79)

Eugen Wüster, Einführung in die Allgemeine Terminologielehre und Terminologische Lexikographie. (hg. L. Bauer). Wien, 1979.

(Zehnder85)

Carl August Zehnder, Informationssysteme und Datenbanken, Zürich 1985 (3).

(Zgusta85)

Ladislav Zgusta(hg), Probleme des Wörterbuchs, Darmstadt, 1985

(Zimmermann72)

Harald Zimmermann, Das Lexikon in der maschinellen Sprachanalyse.
In: Athenäum Skripten. Linguistik 7. Frankfurt a.M., 1972.

(Zimmermann78)

Harald Zimmermann, Probleme der automatischen Indexierung von
Fachtexten am Beispiel juristischer Dokumente. In: S. Koller, P.L.
Reichert, K. Überla, Klartextverarbeitung. Heidelberg, 1978,
p.112.

(Zimmermann83)

Harald Zimmermann, Nutzbarmachung und Nutzung maschineller
Wörterbücher in der Fachinformation und im Büro. In: SDV 6/82,
Tübingen, p.5ff.

(Zoeppritz84)

Magdalena Zoeppritz Syntax for German in the User Speciality
Languages System, Tübingen, 1984

LEBENS LAUF

Eltern	Anna Huonker, geb. Frei Gustav Huonker
19. April 1952	geboren in Zürich
1959 - 1965	Primarschule
1965 - 1971	Realgymnasium
September 1971	B-Matura
1971 - 1978	Studium an der Universität Zürich mit Hauptfach Germanistik
1977 - 1978	Halbtagsstelle als Bibliothekar am Seminar für Allgemeine Sprachwissenschaft der Universität Zürich
Mai 1978	Lizentiat der Philosophischen Fakultät I
1979 - 1980	Auslandaufenthalt in Moskau als Übersetzer beim 'Progress'-Verlag und verschiedenen Firmen (Russisch, Englisch, Deutsch)
1980 - 1982	Programmierer bei der Zentralstelle für EDV der Stadt Zürich
April 1982	Wiederimmatrikulation an der Universität
1982 - 1986	Lehrer an der Berufsschule und an der Kaufmännischen Schule in den Fächern Deutsch und EDV
November 1985	Höheres Lehramt in Deutsch und Philosophie
Oktober 1986	Einreichen der Dissertation bei Professor Harald Burger.

Fragebogen

MASCHINENLESBARE LEXIKA: FRAGEBOGEN

Anlässlich meiner Dissertation befasse ich mich mit der Struktur von maschinellen Lexika.

Im Sinne eines Beitrags zur wissenschaftlichen Kommunikation innerhalb des Bereichs "computergestützte Sprachverarbeitung" möchte ich Sie bitten, folgende Fragen zu beantworten. Eventuell lassen sich bei der Auswertung Ihrer Antworten gewisse Trends zu einer Standardisierung maschineller Lexika im deutschen Sprachbereich feststellen. Ich bin Ihnen sehr zu Dank verpflichtet, wenn Sie sich die Mühe nehmen, diesen Fragebogen auszufüllen. Dadurch ermöglichen Sie mir eine angemessene Darstellung Ihres Projekts in meiner Dissertation. Ohne Ihre Angaben wird eine adäquate Beschreibung und Präsentation Ihres Wörterbuchs stark eingeschränkt.

Ich schreibe meine Dissertation an der Universität Zürich bei Professor H. Burger.
Vielleicht kann ich Ihnen zu einem späteren Zeitpunkt auch einmal mit Auskünften behilflich sein.
Ich bedanke mich für Ihre Hilfe.

FRAGEN:

1. Verwenden Sie mehrere Typen von Lexika?

. . .

☐	Vollformenwörterbuch
☐	Grundformenwörterbuch
☐	Stammwörterbuch

andere

2. Welchen Zwecken dient das Lexikon?

. . .

☐	Lemmatisierung
☐	Syntax
☐	Semantik
☐	maschinelle Uebersetzung
☐	Information, Dokumentation
☐	Künstliche Intelligenz

andere

3. Welches Wortklassenschema wird zugrundegelegt?

- Wortarten (in die Tabelle ev. Abkürzungen eintragen)

- Welche Wortarten unterteilen Sie in welche Unterklassen?

- Welche Wortarten fassen Sie zu Oberklassen zusammen?

4. Welche Informationsfelder werden pro Worteintrag codiert?
(zusammengehörende Informationsfelder (Codes) als solche markieren)

WORTEINTRAG INF.1 INF.2 INF.3 ...

INF.n

Falls die Einträge nach Wortarten differenziert codiert werden, dies in analogen Tabellen darstellen. (Wenn Platz nicht ausreicht, separates Blatt benutzen)

5. Welchen Rechner benutzen Sie?

6. Ist Ihr Verfahren ☐ oder ☐ fileorientiert
datenbankorientiert?

7. Falls fileorientiert:

-Aus welchen Feldern ist ein logischer Record aufgebaut?

Rec.art	Wort	Codes im Detail
---------	------	-----------------

Tabelle für Codes:

-Welche Recordarten benützen Sie?

(falls Platz nicht ausreichend, separates Blatt benutzen)

-Entsprechen Ihre Recordarten den Wortarten? ☐ ja
☐ nein

8. Falls datenbankorientiert:

-Welchen DB-typ
verwenden Sie?

☐
☐
☐
☐

CODASYL
Baumhierarchie
Inverted File
Relational

andere

. . .

-Wie lautet die Bezeichnung Ihrer Datenbank als kommerziell
erwerbbares DBMS?

-Wie lässt sich die Abspeicherung eines Worteintrags in der
Terminologie Ihres Datenbanksystems beschreiben? (Falls es
diese Terminologie zulässt, benutzen Sie für die Antwort
bitte die Tabellen unter Frage Nr. 7 , sonst formulieren
Sie die Antwort hier unter Frage Nr. 8)

-Welche Möglichkeiten vermissen Sie bei Ihrem DBMS?

9. (Dies ist wieder eine allgemeine Frage)
Sind ausser der codierten Lexikoninformation noch andere
Informationen codiert?

☐

Affixlisten

andere

. . .

10. In welcher Programmiersprache operieren Sie über dem Wörter-
buch?

11. mit welchen Zugriffsalgorithmen operieren Sie?

☐
☐
☐

VSAM
HASH
B-Baum

andere

. . .

12. Welche Felder können als Zugriffsschlüssel verwendet werden?

☐
☒
☐

nur ein Schlüssel
alle Felder

. . .

andere

13. Wie gross ist Ihr Wörterbuch?

☐
☒
☒

Anzahl Einträge
Anzahl Grundformen
Anzahl abgeleiteter Formen

. . .

andere Grössenmerkmale

14. Maximale Wortlänge?

15. Maximale Zahl von Informationskategorien (Codes) pro Lexem?

16. Besteht die Möglichkeit Einträge zu verknüpfen?

☐
☒
☒

Verknüpfung Lemma - Formen
Thesaurusrelationen
Frames

. . .

andere

17. Welche LDV-Programme (oder Systeme) benutzen Ihre Wörterbücher?

18. Existiert eine Schnittstelle zur DIN 2341 (z.B. MATER)?

19. Welche konventionellen Wörterbücher dienen als Grundlage?

-

für den Wortschatz
für die Codierung

. . .

andere Informationen

20. Ist es Ihnen möglich, einen typischen Auszug aus dem Wörterbuch und eine Erklärung des Codes beizulegen?

21. Literaturhinweise:

Ich bedanke mich höflich für Ihre Hilfe.
Der Zweck meiner Arbeit besteht darin, dazu beizutragen, den neuesten Stand der Lexikonstrukturierung bei maschinellen Sprachanalyseverfahren in überschaubarer Form darzustellen (deutsches Sprachgebiet).
Ich wäre auch für die Angabe von Adressen von Instituten und Personen dankbar, die sich mit obigen Fragen beschäftigen.
H. Huonker

Ausgefüllte Fragebogen bitte an folgende Adresse senden:
H. Huonker;
Affolternstr. 409
8105 Regensburg
Schweiz

